

Achieving Shrinkage in a Time-Varying Parameter Model Framework*

Angela Bitto and Sylvia Frühwirth-Schnatter[†]

Institute for Statistics and Mathematics, Department of Finance, Accounting and Statistics
WU Vienna University of Economics and Business, Vienna, Austria

November 7, 2016

Abstract

In the present paper, shrinkage for time-varying parameter (TVP) models is investigated within a Bayesian framework, with the aim to automatically reduce time-varying parameters to static ones, if the model is overfitting. This goal is achieved by formulating appropriate shrinkage priors, in particular, for the process variances, based on the normal-gamma prior (Griffin and Brown, 2010). In this way, we extend previous work based on spike-and-slab priors (Frühwirth-Schnatter and Wagner, 2010) and Bayesian Lasso type priors (Belmonte et al., 2014). We develop an efficient MCMC estimation scheme, exploiting boosting ideas such as ancillarity-sufficiency interweaving (Yu and Meng, 2011). Furthermore, we investigate different priors, including the common inverted gamma prior for the process variances, using a predictive analysis and highlight the advantages of using a Kalman mixture approximation to evaluate one-step ahead predictive densities. Our method is applicable both to TVP models for univariate as well as to multivariate time series. This is exemplified through EU area inflation modelling based on the generalized Phillips curve as well as estimating a time-varying covariance matrix based on a TVP Cholesky stochastic volatility model for a multivariate time series of returns derived from the DAX-30 index. Overall, our findings suggest that the family of shrinkage priors introduced in this paper for TVP models is successful in avoiding overfitting, if potentially time-varying parameters are, indeed, static or even insignificant.

Keywords: Bayesian inference; forecasting; hierarchical shrinkage priors; Kalman-filter; lasso; log predictive density scores; normal-gamma prior; state space model.

*We are grateful for many helpful comments on this paper by participants of several ESOBE Meetings (Vienna 2012, Oslo 2013, Gerzensee 2015) and NBER-NSF Time Series Conferences (St. Louis 2014, Vienna 2015, und NYC 2016). We owe special thanks to Mauro Bernardi, discussant of our paper at the 2016 Bolzano Workshop on Forecasting in Finance and Macroeconomics. We acknowledge very helpful comments on preliminary versions of this paper by participants of CFE 2011, the 1st EFaB Workshop (2013), the 7th Rimini Bayesian Econometrics Workshop (2013), the 6th Trondheim Symposium in Statistics (2013), the 1st Bayesian Young Statisticians Meeting (2013), the 1st Vienna Workshop on High Dimensional Time Series in Macroeconomics and Finance (2013), the 4th MCMSki Meeting (2014), and the 12th ISBA World Meeting (2014).

[†]Email-addresses: angela.bitto@wu.ac.at, sfruehwi@wu.ac.at.

1 Introduction

Time-varying parameter (TVP) models are widely used in time series analysis to deal with processes which gradually change over time and provide an interesting alternative to models that allow multiple change points as in Geweke and Jiang (2011). In the present paper, we consider TVP models both with homoscedastic error variances as well as with error variances following a stochastic volatility model (Jacquier et al., 1994). The later model has proven to be useful in various applications, because neglecting time-varying volatilities might lead to overstating the role of time-varying coefficients in explaining structural changes in the dynamics of macroeconomic variables, as exemplified by Sims (2001) and Nakajima (2011). A variety of papers has been dealing with both types of TVP model in recent years. For example, Primiceri (2005) used time-varying structural VARs in a monetary policy application, Dangl and Halling (2012) used TVP models for equity return prediction and Belmonte et al. (2014) used a TVP model to model EU-area inflation.

A big advantage of TVP models is their flexibility in capturing change, however, the risk of overfitting increases as the number of parameters increases. If a considerable number of coefficients is, indeed, constant over the entire observation period, allowing them to change over time leads to considerable loss of statistical inefficiency compared to a model, where these coefficients are assumed to be constant apriori. This will be exemplified in the present paper for a TVP Cholesky SV model in the spirit of Lopes et al. (2015) for a time series of returns from the DAX-30 index, where out of 406 potentially time-varying coefficients only a small fraction actually changes over time.

In the present paper, shrinkage priors for TVP models are investigated within a Bayesian framework, with the aim to automatically reduce time-varying parameters to static ones, if the model is overfitting. This goal is achieved by formulating appropriate shrinkage priors, in particular, for the process variances of the shocks driving the dynamics of the various parameters. There exists a vast literature on introducing shrinkage priors for regression coefficients in a *static* framework, in order to shrink coefficients β_j that are not significant toward zero. One very popular example are spike-and-slab priors which assigns positive probability to the event $\beta_j = 0$ and, thus, can be used for Bayesian variable selection; see O'Hara and Sillanpää (2009) for a recent review. Alternative shrinkage prior specifications in a standard (static) regression model are based on continuous distributions with a pronounced prior spike at zero, one well-known example being the Lasso (Tibshirani, 1996) which can be derived as the Bayesian posterior mode estimator under independent double-exponential priors (Park and Casella, 2008). Further proposed priors, which are all scale mixtures of normals, include the normal-gamma prior (Griffin and Brown, 2010; Caron and Doucet, 2008) and the horseshoe prior (Carvalho et al., 2010), among many others; see Fahrmeir et al. (2010) and Polson and Scott (2011) for a review.

Whereas the literature on shrinkage priors and regularization methods for the static frame-

work has been growing considerably over the recent years, far less work has considered shrinkage priors for the process variances in state space models and TVP models which allow to reduce dynamic coefficients toward static ones, as described above. Frühwirth-Schnatter and Wagner (2011) compare various shrinkage priors for random intercept models, Nakajima and West (2013) present a Bayesian analysis of latent threshold dynamics models and Kalli and Griffin (2014) use a dynamic regression model based on continuous shrinkage priors to forecast equity premium.

To shrink process variances toward zero, Frühwirth-Schnatter and Wagner (2010) use spike-and-slab priors and Belmonte et al. (2014) consider Bayesian Lasso type priors. In the present paper, this goal is achieved by formulating continuous shrinkage priors for process variances based on the normal-gamma prior (Griffin and Brown, 2010; Caron and Doucet, 2008). We show that this prior is more flexible than the Bayesian Lasso prior and yields posterior distributions with a very pronounced prior spike at zero for coefficients which are non time-varying, while at the same time overshrinkage is avoided for time-varying coefficients. An additional shrinkage prior allows to shrink static coefficients to coefficients which are not significant. As a result, we are able to discriminate between time-varying coefficients, coefficients which are significant, but static and insignificant coefficients. We investigate different prior settings, including the common inverted gamma prior for the process variances, using a predictive analysis in the spirit of Geweke and Amisano (2010).

We develop an efficient MCMC estimation scheme to apply our novel approach. To improve MCMC performance, we exploiting boosting ideas such as ancillarity-sufficiency interweaving (Yu and Meng, 2011). A scale-mixture representation of all shrinkage priors allows us to implement full conditional Gibbs sampling, thus avoiding MH steps which are typically needed to implement MCMC methods for non-Gaussian state space models, see e.g. Geweke and Tanizaki (1999).

The rest of the paper is structured as follows. Section 2 specifies the model and discusses in detail our novel shrinkage method in the context of TVP models. In Section 3, we present the Gibbs sampler and discuss the interweaving step. Section 4 addresses predictive evaluation and compares various approximations for the predictive density. Section 5 extends our method to a multivariate framework. Section 6 presents a simulated data example and Section 7 exemplifies our approach through EU area inflation modelling based on the generalized Phillips curve as well as estimating a time-varying covariance matrix based on a TVP Cholesky SV model for a multivariate time series of returns derived from the DAX-30 index. Section 8 concludes.

2 Model specification

2.1 Time-varying parameter models

Starting point for our research is the well known state space model, which has been studied in many fields, see e.g. West and Harrison (1997) and Durbin and Koopman (2000) for a comprehensive review. For the ease of exposition, we will assume in this section that we are modelling a univariate time series y_t , observed for T time periods $t = 1, \dots, T$. Time-varying parameter models for multivariate time series \mathbf{y}_t will be discussed in Section 5. The time series observations y_t are supposed to be driven by latent variables, summarized in a d -dimensional state vector $\boldsymbol{\beta}_t$ which we are unable to observe. The time-varying parameter (TVP) model is a special case of a state space model and can be regarded as a regression model with time-varying regression coefficients $\boldsymbol{\beta}_t$ following a random walk:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}), \quad (1)$$

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \quad (2)$$

where $\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{td})$ is a d -dimensional row vector, containing the regressors of the model, one of them being a constant (e.g. $x_{t1} \equiv 1$). The unknown initial value $\boldsymbol{\beta}_0$ is assumed to follow a normal prior distribution,

$$\boldsymbol{\beta}_0 | \boldsymbol{\beta}, \mathbf{Q} \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{P}_0 \mathbf{Q}), \quad (3)$$

with prior mean $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)'$ being fixed, but unknown regression coefficients and $\mathbf{P}_0 = \text{Diag}(P_{0,11} \dots P_{0,dd})$ being a diagonal matrix. Furthermore, $\boldsymbol{\beta}_0$ is independent of (ε_t) and $(\boldsymbol{\omega}_t)$, which are independent Gaussian white noise processes.

We assume that $\mathbf{Q} = \text{Diag}(\theta_1, \dots, \theta_d)$ is a diagonal matrix,¹ hence each element β_{jt} of $\boldsymbol{\beta}_t = (\beta_{1t}, \dots, \beta_{dt})'$ follows a random walk, i.e. for $j = 1, \dots, d$,

$$\beta_{jt} = \beta_{j,t-1} + \omega_{jt}, \quad \omega_{jt} \sim \mathcal{N}(0, \theta_j), \quad (4)$$

with initial value $\beta_{j0} | \beta_j, \theta_j \sim \mathcal{N}(\beta_j, \theta_j P_{0,jj})$. Hence, θ_j is the process variance governing the dynamics of the time-varying coefficient β_{jt} .

Concerning the error variances in the measurement equation (2), we consider the homoscedastic case, $\sigma_t^2 \equiv \sigma^2$ for all $t = 1, \dots, T$, as well as more flexible model specifications, where σ_t^2 is time-dependent. To capture heteroscedasticity, we use a stochastic volatility (SV) specification

¹Eisenstat et al. (2014) discusses an extension where the error covariance matrix in the state equation is a full matrix instead of a diagonal matrix.

as in Jacquier et al. (1994) where $\sigma_t^2 = e^{h_t}$ and the log volatility h_t follows an AR(1) process:²

$$h_t | h_{t-1}, \mu, \phi, \sigma_\eta^2 \sim \mathcal{N}(\mu + \phi(h_{t-1} - \mu), \sigma_\eta^2). \quad (5)$$

In this setup, the latent time volatility process $\mathbf{h} = (h_0, \dots, h_T)$ is not observed and the initial state h_0 is assumed to follow the stationary distribution of the autoregressive process, i.e. $h_0 | \mu, \phi, \sigma_\eta^2 \sim \mathcal{N}(\mu, \sigma_\eta^2 / (1 - \phi^2))$.

An important building block of our approach is a non-centered parametrization of this TVP model in the spirit of Frühwirth-Schnatter and Wagner (2011) who introduced a non-centered parametrization of the basic structural state space model. First, we define d independent random walk processes $\tilde{\beta}_{jt}$ ($j = 1, \dots, d$) with standard normal independent increments, i.e.

$$\tilde{\beta}_{jt} = \tilde{\beta}_{j,t-1} + \tilde{\omega}_{jt}, \quad \tilde{\omega}_{jt} \sim \mathcal{N}(0, 1), \quad (6)$$

and initial value $\tilde{\beta}_{j0} \sim \mathcal{N}(0, P_{0,jj})$. Then, we rewrite state space model (2) and (4) by combining the d state equations for $\tilde{\beta}_{jt}$ given in (6) with following non-centered observation equation:

$$y_t = \mathbf{x}_t \boldsymbol{\beta} + \mathbf{x}_t \text{Diag}(\sqrt{\theta_1}, \dots, \sqrt{\theta_d}) \tilde{\boldsymbol{\beta}}_t + \varepsilon_t. \quad (7)$$

As can easily be verified, the resulting state space model with state vector $\tilde{\boldsymbol{\beta}}_t = (\tilde{\beta}_{1t}, \dots, \tilde{\beta}_{dt})'$ is a non-centered parametrization of the TVP model (2) and (4), where the non-centered observation equation (7) contains the unknown fixed regression coefficient β_1, \dots, β_d as well as (square roots of) the unknown process variances $\theta_1, \dots, \theta_d$.

Note that the initial value in the non-centered parameterization is assumed to be random (i.e. $\tilde{\boldsymbol{\beta}}_0 \sim \mathcal{N}_d(0, \mathbf{P}_0)$) rather than zero (i.e. $\tilde{\boldsymbol{\beta}}_0 = 0$) as in earlier work (Frühwirth-Schnatter and Wagner, 2011; Belmonte et al., 2014). We found that this additional randomness avoids overshrinking of the time-varying parameters $\boldsymbol{\beta}_t$ toward $\boldsymbol{\beta}$ for the first few time points.

2.2 Bayesian inference

Having set up the time-varying parameter model framework, we perform inference within a Bayesian framework, choosing a new family of prior distributions for the unknown model parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)'$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$. To avoid any scaling issues, we assume that all covariates, except (of course) the intercept, have been standardized such that for each j the average of x_{tj} over t is zero and the sample variance is equal to 1.

The focus of the present paper is to introduce shrinkage, in particular for the process variances θ_j , through a flexible class of prior distributions to be introduced in the Subsection 2.3. The rationale behind this is to pull time-varying regression coefficients $\beta_{j0}, \beta_{j1}, \dots, \beta_{jT}$ toward the

²Alternative specifications are possible such as switching state space models (Frühwirth-Schnatter, 2001).

fixed regression coefficient β_j , if the TVP model is over fitting and the effect of the j th covariate x_{tj} is, in fact, *not* changing over time. This requires the definition of priors on the process variances θ_j that are able to shrink θ_j toward the boundary value 0, but, at the same time, are flexible enough to avoid overshrinking for regression coefficients that are, actually, changing over time t and are characterized by a non-zero process variance ($\theta_j \neq 0$).

If the process variance θ_j is pulled toward zero, then the (practically constant) effect of the covariate x_{tj} is significant, whenever the corresponding fixed regression effect is non-zero ($\beta_j \neq 0$). In high dimensions, where we expect many coefficients to be static, we are also interested in identifying covariates x_{tj} that are insignificant throughout the whole observation period. As these coefficients are characterized by a parameter setting where both $\beta_j = 0$ and $\theta_j = 0$, we also introduce shrinkage priors on the fixed regression coefficients β_j in Subsection 2.4.³

Rather than fixing the variance $P_{0,jj}$ of the initial distribution $\beta_{j0}|\beta_j, \theta_j \sim \mathcal{N}(\beta_j, \theta_j P_{0,jj})$, we found it useful to let $P_{0,jj}$ be an unknown scale parameter, with (conditionally conjugate) prior $P_{0,jj} \sim \mathcal{G}^{-1}(\nu_P, (\nu_P - 1)c_P)$, where the hyper parameters are chosen as $c_P = 1$ and $\nu_P = 20$.

Finally, we employ commonly used priors for the parameters of the error distribution in the observation equation (1). Hence, for the homoscedastic case, we consider the following hierarchical prior:

$$\sigma^2|C_0 \sim \mathcal{G}^{-1}(c_0, C_0), \quad C_0 \sim \mathcal{G}(g_0, G_0), \quad (8)$$

with hyperparameters c_0 , g_0 , and G_0 . In our practical applications, $c_0 = 2.5$, g_0 is a small integer, e.g. $g_0 = 5$, and $G_0 = g_0/E(\sigma^2)(c_0 - 1)$, with $E(\sigma^2)$ being a prior guess of the first moment of σ^2 .

In the SV framework (5), the unknown model parameters are the level $\mu \in \mathbb{R}$, the persistence $\phi \in (-1, 1)$, and the volatility of log variance $\sigma_\eta^2 \in \mathbb{R}^+$. Priors for these parameters are chosen as in Kastner and Frühwirth-Schnatter (2014): $p(\mu, \phi, \sigma_\eta^2) = p(\mu)p(\phi)p(\sigma_\eta^2)$, where $\mu \sim \mathcal{N}(b_\mu, B_\mu)$, the prior of ϕ is chosen according to $(\phi + 1)/2 \sim \mathcal{B}(a_0, b_0)$, and $\sigma_\eta^2 \sim \mathcal{G}(\frac{1}{2}, \frac{1}{2B_\sigma})$, with hyperparameters b_μ , B_μ , a_0 , b_0 , and B_σ . In our practical applications, $b_\mu = 0$, $B_\mu = 100$, $a_0 = 20$, $b_0 = 1.5$, and $B_\sigma = 1$.

2.3 Shrinking process variances through single, double and triple gamma priors

A very popular prior choice for the process variance θ_j is the inverted gamma prior distribution, which is the conjugate prior for θ_j in the centered parameterization, see e.g. Petris et al. (2009):

$$\theta_j|s_0, S_0 \sim \mathcal{G}^{-1}(s_0, S_0). \quad (9)$$

³It should be noted that the data are not informative about β_j , if $\theta_j > 0$. In this case, it is of more interest to consider the initial regression coefficient β_{j0} rather than β_j . For $\theta_j = 0$, β_{j0} and β_j coincide.

However, as has been shown by Frühwirth-Schnatter and Wagner (2011), this prior fails to introduce shrinkage as it is bounded away from zero.

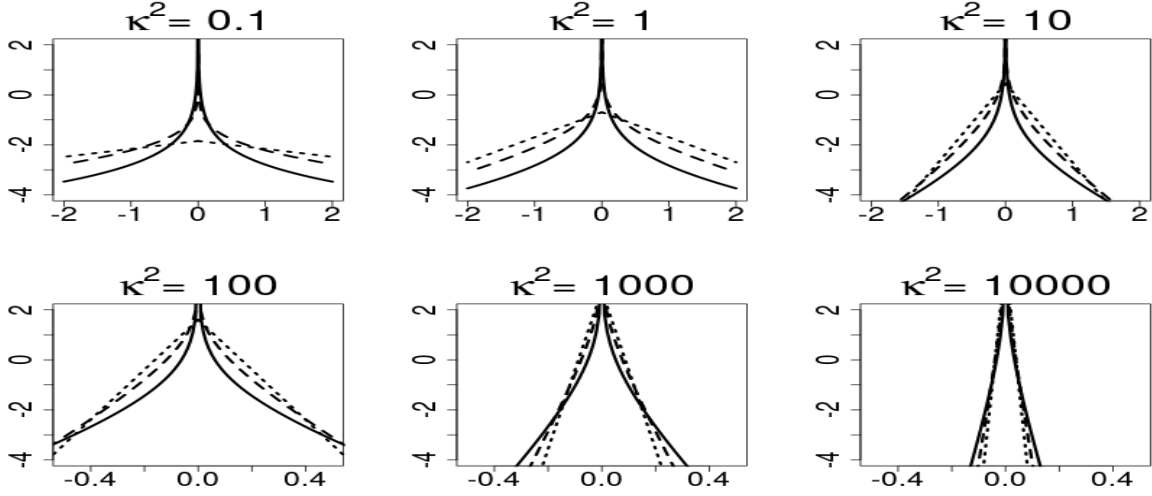


Figure 1: Log of the prior density $\log p(\sqrt{\theta_j}|\kappa^2)$ for different values of κ^2 and $a^\xi = 0.1$ (solid line), $a^\xi = 1/3$ (dashed line) and $a^\xi = 1$ (dotted line).

As noted by Frühwirth-Schnatter (2004), shrinkage priors for process variances are most naturally introduced in the non-centered parameterization (7). It is essential to leave the scale parameter $\sqrt{\theta_j}$, which has a positive and a negative root $\pm\sqrt{\theta_j} \in \mathbb{R}$, unconstrained and to allow $\sqrt{\theta_j}$ to take on positive and negative values. In this way, shrinkage of θ_j toward the boundary value 0 can be achieved by shrinking $\sqrt{\theta_j}$ (which no longer lies on the boundary of the parameter space) toward 0 through an appropriate shrinkage prior. One such possibility was proposed by Frühwirth-Schnatter and Wagner (2011), who realized that the conjugate prior for $\sqrt{\theta_j}$ in the non-centered parameterization (7) is the normal distribution and assumed that $\sqrt{\theta_j}$ is Gaussian with zero prior mean and fixed scale parameters ξ_j^2 :

$$\sqrt{\theta_j}|\xi_j^2 \sim \mathcal{N}(0, \xi_j^2) \quad \Leftrightarrow \quad \theta_j|\xi_j^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_j^2}\right). \quad (10)$$

Evidently, prior (10) substitutes the inverted gamma prior (9) by a “single gamma” prior, namely a χ_1^2 -distribution.⁴

To discriminate between static and random components, Frühwirth-Schnatter and Wagner (2011) introduced spike-and-slab priors, where $\xi_j^2 = 0$ with positive prior probability and ξ_j^2 is fixed (e.g. $\xi_j^2 = 10$), otherwise. Instead of using spike-and-slab priors, Belmonte et al. (2014) extended prior (10) by adding two levels of hierarchy to define a hierarchical Bayesian Lasso shrinkage prior, where ξ_j^2 follows an exponential distribution.

⁴We use the parametrization of the gamma distribution with pdf given by $f_G(y; \alpha, \beta) = \beta^\alpha y^{\alpha-1} e^{-\beta y} / \Gamma(\alpha)$.

In the present paper, we pursue a similar approach, but introduce a more general family of shrinkage priors for state space models derived from the normal-gamma prior, introduced by Griffin and Brown (2010) for variable selection in standard regression models and applied in Caron and Doucet (2008) to multivariate regression models. The main idea is to use the normal-gamma prior as a prior for $\sqrt{\theta}_j$ in the non-centered state space model, extending (10). The normal-gamma prior is a scale mixture of normals (Andrews and Mallows, 1974), with following hierarchical representation:

$$\sqrt{\theta}_j | \xi_j^2 \sim \mathcal{N}(0, \xi_j^2), \quad \xi_j^2 | a^\xi, \kappa^2 \sim \mathcal{G}(a^\xi, a^\xi \kappa^2 / 2). \quad (11)$$

In terms of the process variances, (11) implies that θ_j follows a “double gamma” prior:

$$\theta_j | \xi_j^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_j^2}\right), \quad \xi_j^2 | a^\xi, \kappa^2 \sim \mathcal{G}(a^\xi, a^\xi \kappa^2 / 2). \quad (12)$$

For $a^\xi = 1$, the gamma distribution reduces to the exponential distribution and the Bayesian Lasso prior considered by Belmonte et al. (2014) results as a special case of our prior.

The hierarchical representations (11) will be exploited for MCMC estimation, however, a closed form expression is available both for the marginal density $p(\theta_j | \kappa^2)$ where ξ_j^2 is integrated out as well as for $p(\sqrt{\theta}_j | \kappa^2)$:⁵

$$\begin{aligned} p(\sqrt{\theta}_j | \kappa^2) &= \frac{\sqrt{a^\xi \kappa^2}^{a^\xi + 1/2}}{\sqrt{\pi} 2^{a^\xi - 1/2} \Gamma(a^\xi)} |\sqrt{\theta}_j|^{a^\xi - 1/2} K_{a^\xi - 1/2}(\sqrt{a^\xi \kappa^2} |\sqrt{\theta}_j|), \\ p(\theta_j | \kappa^2) &= \frac{\sqrt{a^\xi \kappa^2}^{a^\xi + 1/2}}{\sqrt{\pi} 2^{a^\xi - 1/2} \Gamma(a^\xi)} (\theta_j)^{a^\xi/2 - 3/4} K_{a^\xi - 1/2}(\sqrt{a^\xi \kappa^2 \theta_j}), \end{aligned} \quad (13)$$

where $K_p(\cdot)$ is the modified Bessel function of the second kind with index p .

The display of the log density $\log p(\sqrt{\theta}_j | \kappa^2)$ for different values of κ^2 and a^ξ in Figure 1 shows that the double gamma prior with $a^\xi \leq 1$ is a typical example of a global-local shrinkage prior (Polson and Scott, 2011). Evidently, a spike at zero occurs for each of these densities, however the amount of mass placed close to zero strongly depends on the choice of a^ξ and κ^2 . From representation (12) we obtain that, marginally, $E(\theta_j) = \frac{2}{\kappa^2}$, whereas

$$V(\theta_j) = E(\theta_j^2) - E(\theta_j)^2 = 3E((\xi_j^2)^2) - \frac{4}{\kappa^4} = \frac{12}{a^\xi \kappa^4} + \frac{8}{\kappa^4} = E(\theta_j)^2 (2 + 3/a^\xi).$$

Hence, independently of a^ξ , the hyperparameter κ^2 controls the global level of shrinkage, which is the stronger, the larger κ^2 . At the same time, also $V(\theta_j)$ decreases, as κ increases. Therefore,

⁵Note that $F_{\theta_j}(c) = \Pr(\theta_j \leq c) = \Pr(-\sqrt{c} \leq \sqrt{\theta}_j \leq \sqrt{c}) = 2F_{\sqrt{\theta}_j}(\sqrt{c})$, where $F_{\theta_j}(\cdot)$ is the cdf of $\sqrt{\theta}_j$. Therefore, $p(\theta_j | \kappa^2) = p(\sqrt{\theta}_j | \kappa^2) / \sqrt{\theta}_j$.

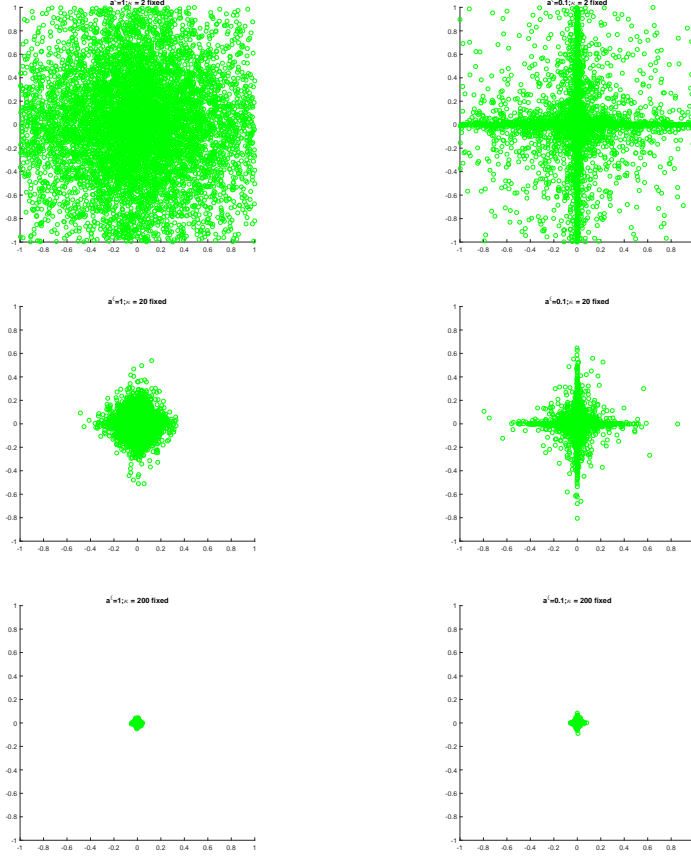


Figure 2: Simulations from the double Gamma prior $p(\sqrt{\theta}_1, \sqrt{\theta}_2 | \kappa^2)$ for $a^\xi = 1$ (left hand side) and $a^\xi = 0.1$ (right hand side) and different value of κ^2 (top: $\kappa^2 = 2$, middle: $\kappa^2 = 20$, bottom: $\kappa^2 = 200$).

the larger κ^2 , the more mass is placed close to zero. It is also evident from Figure 1 that very large values for κ^2 lead to extreme shrinkage.

On the other hand, the term $3/a^\xi$ – which is also equal to the excess kurtosis of $\sqrt{\theta}_j$ – controls local adaption to the global level of shrinkage, with more local adaption, the smaller a^ξ . As a^ξ decreases, the excess kurtosis of $\sqrt{\theta}_j$ increases and the tails of $p(\sqrt{\theta}_j | \kappa^2)$ become thicker, in particular for values of κ that are not too large, see again Figure 1. Therefore, the shape parameter a^ξ is of great importance for introducing adaptive sparsity in TVP models and needs to be chosen carefully. We will work with values equal to and smaller than 1 and reduce this hyperparameter down to 0.05.

It is also illuminating to investigate the joint marginal prior distribution of $(\theta_1, \dots, \theta_d)$ or (equivalently) of $(\sqrt{\theta}_1, \dots, \sqrt{\theta}_d)$ given κ^2 . Since the random prior variances ξ_j^2 in (12) are

drawn independently, regardless of the choice of a^ξ , also marginally the double gamma prior is characterized by prior conditional independence of $(\theta_1, \dots, \theta_d)$ given a fixed value of κ^2 :

$$p(\theta_1, \dots, \theta_d | \kappa^2) = \prod_{j=1}^d p(\theta_j | \kappa^2).$$

For illustration, Figure 2 shows simulations from the joint prior $p(\sqrt{\theta_1}, \sqrt{\theta_2} | \kappa^2)$ for $d = 2$ for various values of a^ξ and κ^2 . Not surprisingly from the previous discussions, for the same value of κ^2 , the double gamma with $a^\xi = 0.1$ has a pronounced spike at 0 with fat tails in both directions of $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ and provides more flexible shrinkage compared to the Bayesian Lasso prior with $a^\xi = 1$. For the Bayesian Lasso, large values of κ^2 (e.g. $\kappa^2 = 200$) are needed to introduce strong shrinkage toward 0.

To infer an appropriate value of κ^2 from the data, yet another layer of hierarchy is added, by assuming that the hyperparameter κ^2 is random, following again a gamma distribution:

$$\kappa^2 \sim \mathcal{G}(d_1, d_2). \quad (14)$$

The corresponding hierarchical prior, combining (12) with (14), is a “triple gamma” prior. Apart from adapting the hyperparameter κ^2 to the time series data at hand, prior (14) also introduces prior dependence among $(\theta_1, \dots, \theta_d)$, given d_1 and d_2 . Such prior dependence is desirable in particular in a shrinkage environment, where only few variances are expected to be different from 0. In this case, whether a certain process variance is shrunk toward 0 also depends on how many of the other process variances are shrunk toward 0.

For illustration, Figure 3 shows simulations from the joint prior $p(\sqrt{\theta_1}, \sqrt{\theta_2} | d_1, d_2)$ for $d = 2$ for various values of $d_1 = d_2$. Again, for the same set of hyperparameters, the prior with $a^\xi = 0.1$ shows a much more flexible shrinkage behaviour than the prior with $a^\xi = 1$ which corresponds to the hierarchical Bayesian lasso prior consider by Belmonte et al. (2014).

2.4 Shrinkage priors for the fixed regression coefficients

If θ_j is shrunk toward 0, then this pulls β_{j0} and all subsequent values β_{jt} toward the fixed regression parameter β_j . For this case, it is also relevant to allow shrinkage of β_j toward 0. Once more, the normal-gamma prior is employed as a shrinkage prior for β_j :

$$\beta_j | \tau_j^2 \sim \mathcal{N}(0, \tau_j^2), \quad \tau_j^2 | a^\tau, \lambda^2 \sim \mathcal{G}(a^\tau, a^\tau \lambda^2 / 2). \quad (15)$$

Again, for $a^\tau = 1$, prior (15) reduces to the prior discussed in Belmonte et al. (2014).

A closed form expression, comparable to (13), is available for $p(\beta_j | \lambda^2)$, with expectation $E(|\beta_j|) = \sqrt{\frac{4}{\pi a^\tau \lambda^2}} \frac{\Gamma(a^\tau + 1/2)}{\Gamma(a^\tau)}$, $V(\beta_j) = \frac{2}{\lambda^2}$, while the excess kurtosis is given by $\frac{3}{a^\tau}$. Also, another

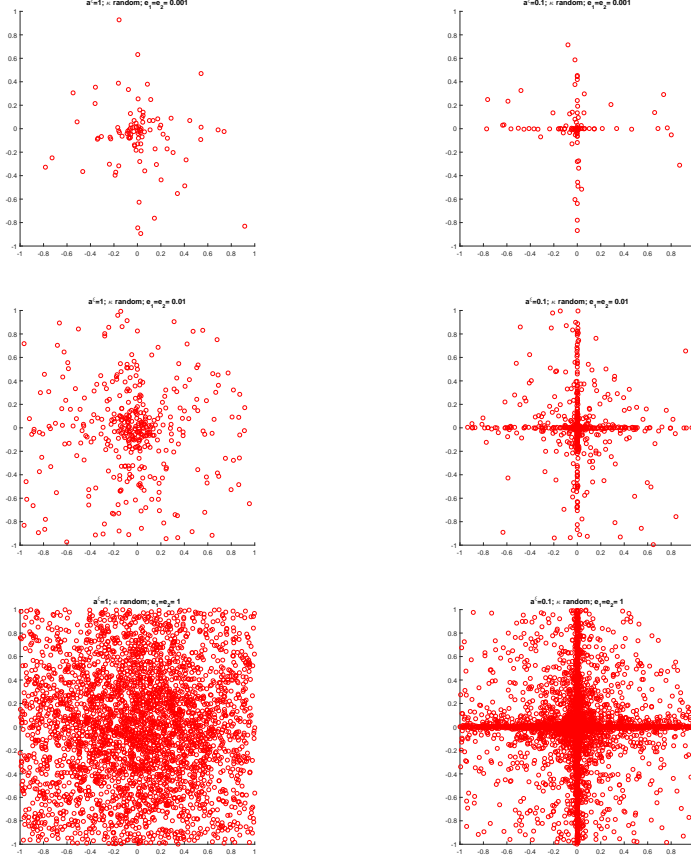


Figure 3: Simulations from the triple gamma prior $p(\sqrt{\theta_1}, \sqrt{\theta_2}|d_1, d_2)$ for $a^\xi = 1$ (left hand side) and $a^\xi = 0.1$ (right hand side) and different value of d_1 and d_2 (top: $d_1 = d_2 = 0.001$, middle: $d_1 = d_2 = 0.01$, bottom: $d_1 = d_2 = 1$).

layer of hierarchy is added, by assuming that λ^2 is random:

$$\lambda^2 \sim \mathcal{G}(e_1, e_2), \quad (16)$$

with similar considerations concerning the choice of a^τ and e_1 and e_2 as before.

3 MCMC Estimation

To carry out Bayesian inference under the shrinkage priors introduced in the previous section, we develop efficient schemes for full conditional Markov chain Monte Carlo (MCMC) sampling. We assume that all hyperparameters in the various priors have been selected, i.e. $a^\tau, a^\xi, d_1, d_2, e_1, e_2$

in the priors for β and \mathbf{Q} , c_P, ν_P in the prior of $P_{0,11}, \dots, P_{0,dd}$, as well as c_0, C_0, G_0 for homoscedastic variances and $a_0, b_0, B_\mu, B_\sigma$ for heteroscedastic variances following an SV model.

Bayesian inference operates in the latent variable formulation of the time-varying parameter model and, as usual for these type of models, relies on data augmentation. Depending on the parametrization, either $\beta = (\beta_0, \beta_1, \dots, \beta_T)$ (for the centered parameterization) or $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_T)$ (for the non-centered parameterization) is considered as latent variable. For the stochastic volatility specification for the error variance given in (5), the latent log volatilities $\mathbf{h} = (h_0, \dots, h_T)$ are introduced as an additional set of latent variables.

For the centered parameterization under the common inverted gamma prior for the process variances θ_j introduced earlier in (9), the steps in the corresponding Gibbs sampler are totally standard, see e.g. Frühwirth-Schnatter (2006, Chapter 13). However, as discussed above, this standard MCMC scheme does not introduce appropriate shrinkage on the process variances. Furthermore, like many MCMC schemes which alternate between sampling from the full conditionals of the latent states and the model parameters, this strategy is known to yields inefficient Gibbs samplers. In particular, if some of the process variances are close to 0, slow convergence and poor mixing (i.e. high correlation of the posterior draws) can become a serious issue, see e.g. Frühwirth-Schnatter (2004). This may cause enormous autocorrelation of posterior draws – even despite considerable thinning (i.e. keeping only every n th posterior draws) – and can render MCMC inference more or less useless, as badly mixing samplers frequently lead to unreliable estimates.

A large literature has emerged discussing various techniques to improve such algorithms. Reparameterization, for instance, relies on data augmentation in a different parameterization of the model with alternative latent variables. Frühwirth-Schnatter (2004) discusses the relationship between various parametrizations of a state space model and the computational efficiency of the resulting MCMC sampler, see also Papaspiliopoulos et al. (2007). As shown by Frühwirth-Schnatter and Wagner (2010), the non-centered parameterization of a state space model, where unknown model parameters are moved from the latent state equation to the observation equation, proves to be useful, in particular if process variances are close to 0. However, MCMC estimation based on different parameterizations will often be efficient in separate regions of the parameter space, as demonstrated e.g. by Kastner and Frühwirth-Schnatter (2014) in the context of univariate SV models.

We encountered poor mixing behaviour also for TVP models, both with the centered parameterization which is preferable for components that are actually time-varying as well as with the non-centered parametrization which is preferable for (nearly) constant components. For practical time series analysis, both types of coefficients are likely to be present and choosing a computationally efficient parametrization in advance is not possible. We show in Subsection 3.1 how different data augmentation schemes can be combined through a strategy introduced by Yu and Meng (2011) to obtain an efficient sampler, thereby combining “best of both worlds”.

3.1 Efficient full conditional Gibbs sampling for sparse TVP models

In this section, the *ancillarity-sufficiency interweaving strategy* (ASIS) introduced by Yu and Meng (2011) is applied to design a posterior simulation method for TVP models which (a) allows Gibbs sampling under the hierarchical shrinkage priors introduced earlier in Section 2.3, and (b) increases posterior sampling efficiency considerably compared to conventional Gibbs sampling either in the centered or the non-centered parameterization. As we work in a conjugate prior scenario, this allows to set up a straightforward scheme for full conditional MCMC sampling as outlined in Algorithm 1 below. For this scheme, the interweaving strategy turns out to be instrumental for an efficient implementation of the shrinkage priors introduced in this paper.

ASIS provides a principled way of interweaving two different data augmentation schemes by re-sampling certain parameters conditional on the latent variables in the alternative parameterization of the model. This strategy has been successfully employed in a variety of contexts such as univariate SV models (Kastner and Frühwirth-Schnatter, 2014), multivariate factor SV models (Kastner et al., 2016) and dynamic linear state space models (Simpson et al., 2015). In the present paper, boosting MCMC in Algorithm 1 is based on interweaving the centered and the non-centered parameterization of the TVP model introduced in Subsection 2.1. More specifically, we use the non-centered parametrization as baseline, and interweave into the centered parameterization through Step (b*). More details on this step are given in Section 3.2.

Algorithm 1. Choose starting values for $\beta, \mathbf{Q}, \boldsymbol{\tau} = (\tau_1, \dots, \tau_d)$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d)$, κ^2, λ^2 , and \mathbf{P}_0 , as well as for σ^2 and C_0 (for homoscedastic variances) and repeat the following steps:

- (a) Sample all latent states $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0, \dots, \tilde{\boldsymbol{\beta}}_T)$ in the non-centered parametrization jointly from the multivariate Gaussian distribution $\tilde{\boldsymbol{\beta}}|\boldsymbol{\beta}, \mathbf{Q}, \sigma^2, \mathbf{P}_0 \sim \mathcal{N}_{(T+1) \cdot d}(\boldsymbol{\Omega}^{-1}\mathbf{c}, \boldsymbol{\Omega}^{-1})$ given in (A.1).
- (b) Joint sampling of $\boldsymbol{\alpha} = (\beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d})'$ from the multivariate Gaussian posterior $p(\boldsymbol{\alpha}|\tilde{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\xi}, \sigma^2, \mathbf{y})$ given in (A.3).
- (b*) Boosting through ASIS: for each $j = 1, \dots, d$, redraw the constant coefficient β_j and the square root of the process variance $\sqrt{\theta_j}$ through interweaving into the state equation of the centered parameterization, see Algorithm 2 in Subsection 3.2.
- (c) Sample the prior variances $\xi_j|\theta_j, a^\xi, \kappa^2$ and $\tau_j|\beta_j, a^\tau, \lambda^2$, for $j = 1, \dots, d$, from conditionally independent generalized inverse Gaussian distributions given in (A.5) and (A.6), respectively, and update the hyperparameters $\lambda^2|a^\tau, e_1, e_2, \boldsymbol{\tau}$ and $\kappa^2|a^\xi, d_1, d_2, \boldsymbol{\xi}$ from the gamma distributions given in (A.9) and (A.10).
- (d) Sample (for the homoscedastic model) the error variance $\sigma^2|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, C_0$ from the inverted gamma posterior (18) and the hyperparameter $C_0|\sigma^2$ from the gamma posterior (19).

(e) *Sample the scale parameters of the initial distribution, for each $j = 1, \dots, d$, from $P_{0,j}|\tilde{\beta}_{j0} \sim \mathcal{G}^{-1}\left(\nu_P + \frac{1}{2}, (\nu_P - 1)c_P + \frac{1}{2}\tilde{\beta}_{j0}^2\right)$.*

After discarding a certain amount of initial draws (the *burn-in*), the full conditional sampler iterating Steps (a) to (e) of Algorithm 1 yields draws from the joint posterior distribution $p(\tilde{\beta}, \beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d}, \xi, \tau, \lambda^2, \kappa^2, \sigma^2, C_0, \mathbf{P}_0|\mathbf{y})$ under the hierarchical shrinkage priors outlined in Subsection 2.3.

In Step (a), we sample the latent states $\tilde{\beta} = (\tilde{\beta}_0, \dots, \tilde{\beta}_T)$ in the non-centered parametrization conditional on known parameters $\beta, \mathbf{Q}, \mathbf{P}_0$ and known error variances $\sigma_1^2, \dots, \sigma_T^2$. As an alternative to the commonly used *Forward Filtering Backward Sampling* (FFBS) algorithm (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994), we implemented a multi-move sampling algorithm in the spirit of McCausland et al. (2011) which allows to sample the entire state process $\tilde{\beta}$ *all without a loop* (AWOL; Kastner and Frühwirth-Schnatter (2014)). Full details are provided in Appendix A.1.1.1.

In Step (b), conditional on the latent states $\tilde{\beta}$, a multivariate regression model results from the observation equation (7) of the non-centered state space model. Based on the Gaussian priors appearing in the hierarchical representations of the shrinkage priors (10) and (15), we sample the parameters β_1, \dots, β_d and $\sqrt{\theta_1}, \dots, \sqrt{\theta_d}$ jointly from the conditionally Gaussian posterior given in (A.3); see Appendix A.1.1.2 for details. One major advantage of working with the square root of the process variance $\sqrt{\theta_j}$ instead of θ_j , is that we avoid boundary space problems for small variances, resulting in better mixing behaviour for models with small variances. Due to the symmetry of the likelihood around $\sqrt{\theta_j} = 0$, the marginal posterior density of $\sqrt{\theta_j}$ exhibits a bimodal structure for those models where the true $\sqrt{\theta_j}$ is different from zero. A unimodal posterior located close to zero results, if the true value $\sqrt{\theta_j}$ is close to zero.

Sampling the latent prior variances ξ_j^2 and τ_j^2 in the hierarchical representations of the shrinkage priors (10) and (15) for $\sqrt{\theta_j}$ and β_j in Step (c) is less standard than the other sampling steps. Subsequently, we briefly discuss sampling ξ_j^2 and κ^2 . Similar results are obtained for $\tau_j^2|a^\tau, \lambda^2, \beta_j$ and $\lambda^2|a^\tau, e_1, e_2, \tau$, see Appendix A.1.1.3 for full details. The conditionally normal prior $\sqrt{\theta_j}|\xi_j^2$ in the hierarchical representations given in (10) leads to a likelihood for ξ_j^2 which is the kernel of an inverted gamma density. In combination with the gamma prior $\xi_j^2|a^\xi, \kappa^2 \sim \mathcal{G}(a^\xi, a^\xi\kappa^2/2)$, this leads to a posterior distribution for the prior variance $\xi_j^2|\theta_j, a^\xi, \kappa^2$ which arises from a generalized inverse Gaussian (GIG) distribution:

$$\xi_j^2|\theta_j, a^\xi, \kappa^2 \sim \text{GIG}\left(a^\xi - 1/2, a^\xi\kappa^2, \theta_j\right).$$

To sample from the GIG distribution, we use a method recently proposed by Hörmann and Leydold (2014) which is also implemented in the R-package GIGrv (Leydold and Hörmann, 2011; Hörmann and Leydold, 2015). This method is especially reliable for time-varying parameter models where the second scale parameter of the GIG (i.e. θ_j) can be extremely small and other

sampler tend to fail.

Given the prior variances $\boldsymbol{\xi} = (\xi_1^2, \dots, \xi_d^2)$, we sample κ^2 from the conditional gamma posterior:

$$\kappa^2 | a^\xi, d_1, d_2, \boldsymbol{\xi} \sim \mathcal{G} \left(d_1 + a^\xi d, d_2 + \frac{\bar{\xi}^2}{2} a^\xi d \right), \quad (17)$$

where $\bar{\xi}^2 = \frac{1}{d} \sum_{j=1}^d \xi_j^2$ is the average prior variance in the shrinkage prior for $\sqrt{\theta_j}$. Hence, if $\bar{\xi}^2$ is small, i.e. a lot of sparsity is present, then the posterior expectation of κ^2 will be proportional to $1/d_2$, implying large posterior draws for κ^2 .

Step (d) depends on modelling assumptions concerning the error variance σ_t^2 . For the homoscedastic error specification based on the inverted gamma prior (8), the posterior of $\sigma^2 | \mathbf{y}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, C_0$ is again an inverted gamma distribution:

$$\sigma^2 | \mathbf{y}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, C_0 \sim \mathcal{G}^{-1} \left(c_0 + \frac{T}{2}, C_0 + \frac{1}{2} \sum_{t=1}^T (y_t - \mathbf{z}_t \boldsymbol{\alpha})^2 \right). \quad (18)$$

Due to the hierarchical structure of prior (8), it is necessary to add an additional sampling step where the prior scale $C_0 | \sigma^2$ is sampled conditional on σ^2 from the conditional gamma posterior at each sweep of the sampler:

$$C_0 | \sigma^2 \sim \mathcal{G} \left(g_0 + c_0, G_0 + \frac{1}{\sigma^2} \right). \quad (19)$$

To implement Step (d) for a time dependent error variance σ_t^2 following the SV model defined in (5), we make use of the efficient MCMC estimation scheme proposed by Kastner and Frühwirth-Schnatter (2014) in which an interweaving strategy for boosting the MCMC estimation of stochastic volatility models has been presented. Furthermore, this MCMC estimation scheme allows for easy incorporation into Algorithm 1 using the R package `stochvol`. Further details can be found in Kastner (2016).

3.2 Boosting Full Conditional MCMC through Interweaving

In Step (b*), we temporarily move from the non-centered to the centered parameterization of the TVP model to resample the constant coefficient β_j and the process variance θ_j . As this interweaving step is designed only to boost MCMC, we need to ensure that the chosen prior distribution is preserved, meaning that the posterior distributions obtained from Algorithm 1 and a sampler without interweaving (i.e. Algorithm 1 without Step (b*)) are identical. To this aim, we need to match the priors between the centered and the non-centered parametrization. Whereas the Gaussian prior $\beta_j | \tau_j^2 \sim \mathcal{N} \left(0, \tau_j^2 \right)$ for the initial value β_j is the same for both

parameterizations, we need to transform the Gaussian prior $\sqrt{\theta_j}|\xi_j^2 \sim \mathcal{N}(0, \xi_j^2)$ to the corresponding prior for the variance θ_j in the centered state equation (4) which yields the gamma prior $\theta_j|\xi_j^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_j^2}\right)$, as discussed in Subsection 2.3. Based on these priors, Step (b*) consists of three parts, outlined in Algorithm 2.

Algorithm 2 (ASIS for TVP models). *For each $j = 1, \dots, d$, redraw the constant coefficient β_j and the square root of the process variance $\sqrt{\theta_j}$ through interweaving into the state equation of the centered parameterization:*

(b*-1) *First, use transformation (20) to move the current posterior draws of the latent process $\tilde{\beta}_{j0}, \dots, \tilde{\beta}_{jT}$ from the non-centered parameterization (6) and (7) to the centered parameterization (1) and (2):*

$$\beta_{jt} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{jt}, \quad t = 0, \dots, T. \quad (20)$$

Furthermore, store the sign of $\sqrt{\theta_j}$.

(b*-2) *Update β_j and $\sqrt{\theta_j}$, conditional on the state process $\beta_{j0}, \dots, \beta_{jT}$ in the centered parameterization:*

(b*-2a) *Draw θ_j^{new} from the conditional posterior $\theta_j|\beta_j, \beta_{j0}, \dots, \beta_{jT}, \xi_j^2, P_{0,jj}$, given by the generalized inverse Gaussian distribution (21).*

(b*-2b) *Draw β_j^{new} from the conditional posterior $\beta_j|\beta_{j0}, \theta_j^{\text{new}}, P_{0,jj}, \tau_j^2$, given by the normal distribution (22).*

(b*-2c) *Calculate the square root $\sqrt{\theta_j^{\text{new}}}$ of the variance and use exactly the same (stored) sign as the old value $\sqrt{\theta_j}$.*

(b*-3) *Based on $\sqrt{\theta_j^{\text{new}}}$ and β_j^{new} , the state process $\tilde{\beta}_{jt}$ in the non-centered parameterization is updated in a deterministic manner through the inverse of transformation (20):*

$$\tilde{\beta}_{jt}^{\text{new}} = (\beta_{jt} - \beta_j^{\text{new}})/\sqrt{\theta_j^{\text{new}}}, \quad t = 0, \dots, T.$$

In Step (b*-2), the posteriors of θ_j and β_j in the centered parameterization, conditional on knowing the state process $\beta_{j0}, \dots, \beta_{jT}$, are easily obtained. First, the conditional posterior

$$p(\theta_j|\beta_j, \beta_{j0}, \dots, \beta_{jT}, \xi_j^2, P_{0,jj}) \propto p(\theta_j)p(\beta_{j0}|\beta_j, \theta_j, P_{0,jj}) \prod_{t=1}^T p(\beta_{jt}|\theta_j, \beta_{j,t-1}),$$

where $\beta_{j0}|\beta_j, \theta_j, P_{0,jj} \sim \mathcal{N}(\beta_j, \theta_j P_{0,jj})$ and $\beta_{jt}|\theta_j, \beta_{j,t-1} \sim \mathcal{N}(\beta_{j,t-1}, \theta_j)$, is the density of a

generalized inverse Gaussian distribution with the following parameters:

$$\theta_j | \beta_j, \beta_{j0}, \dots, \beta_{jT}, \xi_j^2, P_{0,jj} \sim \mathcal{GIG} \left(-\frac{T}{2}, \frac{1}{\xi_j^2}, \sum_{t=1}^T (\beta_{jt} - \beta_{j,t-1})^2 + \frac{(\beta_{j0} - \beta_j)^2}{P_{0,jj}} \right). \quad (21)$$

Note that sampling the process variance θ_j from this GIG posterior deviates from the usual MCMC inference for the centered state space model, since the conditionally conjugate inverted gamma prior (9) is substituted by a prior from the gamma distribution.

Finally, the posterior $p(\beta_j | \beta_{j0}, \theta_j, P_{0,jj}, \tau_j^2)$ is a Gaussian distribution, obtained by combining the prior $\beta_j | \tau_j^2 \sim \mathcal{N}(0, \tau_j^2)$ with the conditional likelihood $\beta_{j0} | \beta_j, \theta_j, P_{0,jj}, \tau_j^2 \sim \mathcal{N}(\beta_j, \theta_j P_{0,jj})$:

$$\begin{aligned} \beta_j | \beta_{j0}, \theta_j, P_{0,jj}, \tau_j^2 &\propto \frac{1}{\sqrt{\theta_j P_{0,jj}}} \exp \left(-\frac{(\beta_{j0} - \beta_j)^2}{2\theta_j P_{0,jj}} \right) \frac{1}{\sqrt{\tau_j^2}} \exp \left(-\frac{\beta_j^2}{2\tau_j^2} \right), \\ &\propto \mathcal{N} \left(\frac{\beta_{j0}\tau_j^2}{\tau_j^2 + \theta_j P_{0,jj}}, \frac{\tau_j^2 \theta_j P_{0,jj}}{\tau_j^2 + \theta_j P_{0,jj}} \right). \end{aligned} \quad (22)$$

4 Predictive Analysis

When working with time series, the ability to forecast and the quality of those predictions are relevant, especially for practitioners. Keeping in mind that in general the value of a future realization of y_{t+1} is uncertain, predictions in a Bayesian context aim at deriving the predictive posterior distribution $p(y_{t+1} | \mathbf{y}^t)$ of the random variable y_{t+1} given observations $\mathbf{y}^t = (y_1, \dots, y_t)$, see Geweke and Amisano (2010) for an excellent review.

In the present paper, we are particularly interested in how the choice of the prior on the process variances θ_j and, also, on the fixed parameter β_j impacts the quality of those predictions. We evaluate the predictive performance of the different shrinkage priors by comparing log-predictive density scores, see Subsection 4.1. Subsection 4.2 exploits how the functional value of $p(y_{t+1} | \mathbf{y}^t)$ can be approximated efficiently, given the observed realization of y_{t+1} .

4.1 Evaluating shrinkage priors through log predictive density scores

Log predictive density score (LPDS) are an often used scoring rule proposed by Good (1952) to compare models; see also Diebold et al. (1998); Diks et al. (2011); Gneiting (2011); Gneiting and Ranjan (2011). Geweke and Keane (2007) and Villani et al. (2009) introduce LPDS for model comparison of econometric models. In the present paper, we use log predictive density scores as a mean to evaluate and compare different shrinkage priors.

As common in this framework, the first t_0 time series observations $\mathbf{y}^{\text{tr}} = (y_1, \dots, y_{t_0})$ are used as a “training sample”, while evaluation is performed for the remaining time observations

y_{t_0+1}, \dots, y_T , by considering the log predictive density,

$$\text{LPDS} = \log p(y_{t_0+1}, \dots, y_T | \mathbf{y}^{\text{tr}}) = \sum_{t=t_0+1}^T \log p(y_t | y^{t-1}) = \sum_{t=t_0+1}^T \text{LPDS}_t^*, \quad (23)$$

where $\log p(y_t | y^{t-1})$ is the one-step ahead log predictive density for time t . The (individual) log predictive density scores $\text{LPDS}_t^* = \log p(y_t | y^{t-1})$ provide a tool to analyse performance separately for each observation y_t , whereas LPDS is an aggregated measure of performance for the entire time series.

As shown by Frühwirth-Schnatter (1995) in the context of selecting time-varying and fixed components for a basic structural state space models, the LPDS defined in (23) has an additional interpretation as log marginal likelihood based on the training sample prior $p(\boldsymbol{\vartheta} | \mathbf{y}^{\text{tr}})$, since

$$p(y_{t_0+1}, \dots, y_T | \mathbf{y}^{\text{tr}}) = \int p(y_{t_0+1}, \dots, y_T | \mathbf{y}^{\text{tr}}, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta} | \mathbf{y}^{\text{tr}}) d\boldsymbol{\vartheta},$$

where $\boldsymbol{\vartheta}$ summarises the unknown model parameters, e.g. $\boldsymbol{\vartheta} = (\beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d}, \sigma^2)$ for the homoscedastic state space model. This provides a sound and coherent foundation for using the log predictive density score LPDS for model – or, in our context, rather prior – comparison.

4.2 Approximating the one-step ahead predictive density

To approximate the predictive density, we use Gaussian sum approximations, which are derived from the draws $(\boldsymbol{\vartheta}^{(m)}, m = 1, \dots, M)$ from the posterior distribution $p(\boldsymbol{\vartheta} | \mathbf{y}^t)$ of the MCMC chain given information up to \mathbf{y}^t , i.e:

$$\begin{aligned} \text{LPDS}_{t+1}^* &= \log p(y_{t+1} | \mathbf{y}^t) = \log \int p(y_{t+1} | \mathbf{y}^t, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta} | \mathbf{y}^t) d\boldsymbol{\vartheta} \\ &\approx \log \left(\frac{1}{M} \sum_{m=1}^M p(y_{t+1} | \mathbf{y}^t, \boldsymbol{\vartheta}^{(m)}) \right), \end{aligned}$$

where the one step ahead predictive density $p(y_{t+1} | \mathbf{y}^t, \boldsymbol{\vartheta})$ is Gaussian conditional on knowing $\boldsymbol{\vartheta}$.

In particular, we derive an approximation which we call the *conditionally optimal Kalman mixture approximation* and which exploits the fact, that the model we are dealing with is conditionally Gaussian given $\boldsymbol{\vartheta} = (\beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d}, \sigma_{t+1}^2)$. Hence, we use the conditional Kalman filter outlined in Appendix A.1.2, to determine the *exact* predictive density $p(y_{t+1} | \mathbf{y}^t, \boldsymbol{\vartheta})$. For each draw $\boldsymbol{\vartheta}^{(m)} = (\beta_1^{(m)}, \dots, \beta_d^{(m)}, \sqrt{\theta_1^{(m)}}, \dots, \sqrt{\theta_d^{(m)}}, \sigma_{t+1}^{2(m)})$ from the posterior $p(\boldsymbol{\vartheta} | \mathbf{y}^t)$, the conditional predictive density $p(y_{t+1} | \mathbf{y}^t, \boldsymbol{\vartheta}^{(m)})$ is a normal distribution,

$$y_{t+1} | \mathbf{y}^t, \boldsymbol{\vartheta}^{(m)} \sim \mathcal{N}_d \left(\hat{y}_{t+1}^{(m)}, S_{t+1}^{(m)} \right),$$

where the corresponding moments $\hat{y}_{t+1}^{(m)}$ and $S_{t+1}^{(m)}$ are easily available from the prediction step of the Kalman filter, based on the filtering density $\tilde{\beta}_t | \mathbf{y}^t, \boldsymbol{\vartheta}^{(m)} \sim \mathcal{N}(\mathbf{m}_t^{(m)}, \mathbf{C}_t^{(m)})$:

$$\begin{aligned}\hat{y}_{t+1}^{(m)} &= \mathbf{x}_{t+1} \boldsymbol{\beta}^{(m)} + \mathbf{F}_{t+1}^{(m)} \mathbf{m}_t^{(m)}, \\ S_{t+1}^{(m)} &= \mathbf{F}_{t+1}^{(m)} (\mathbf{C}_t^{(m)} + \mathbf{I}) \mathbf{F}_{t+1}'^{(m)} + \sigma_{t+1}^{2(m)},\end{aligned}$$

where $\mathbf{F}_{t+1}^{(m)} = \mathbf{x}_{t+1} \text{Diag}(\sqrt{\theta_1^{(m)}}, \dots, \sqrt{\theta_d^{(m)}})$ and \mathbf{I} is the identity matrix.

Given the assumption, that the future regressors \mathbf{x}_{t+1} are known, this yields the following Gaussian mixture approximation for $p(y_{t+1} | \mathbf{y}^t)$:

$$p(y_{t+1} | \mathbf{y}^t) \approx \frac{1}{M} \sum_{m=1}^M f_N(y_{t+1}; \hat{y}_{t+1}^{(m)}, S_{t+1}^{(m)}). \quad (24)$$

Draws from the posterior distribution $p(\boldsymbol{\vartheta} | \mathbf{y}^t)$ are obtained by running the Gibbs sampler outlined in Algorithm 1 for the reduced sample $\mathbf{y}^t = (y_1, y_2, \dots, y_t)$. For a homoscedastic error specification, $\boldsymbol{\vartheta}^{(m)} = (\beta_1^{(m)}, \dots, \beta_d^{(m)}, \sqrt{\theta_1^{(m)}}, \dots, \sqrt{\theta_d^{(m)}}, \sigma^{2(m)})$. For the stochastic volatility specification, $\sigma_{t+1}^{2(m)}$ has to be forecasted first, before the conditional Kalman filter can be applied. To this aim, we exploit the SV model given in (5). For each posterior draw $h_t^{(m)}$, obtained from Algorithm 1, we simulate $h_{t+1}^{(m)}$ from a conditional normal distribution with mean $\mu^{(m)} + \phi^{(m)}(h_t^{(m)} - \mu^{(m)})$ and variance $\sigma_\eta^{2(m)}$ and define $\sigma_{t+1}^{2(m)} = e^{h_{t+1}^{(m)}}$.

As the conditionally optimal Kalman mixture approximation performs *exact* analytical integration with respect to the entire state process $\tilde{\beta}_0, \dots, \tilde{\beta}_{t+1}$, not surprisingly, we found that this method outperforms alternative approximations. Belmonte et al. (2014), for instance, employ a purely simulation-based approach to approximate $p(y_{t+1} | \mathbf{y}^t)$. Based on the same output of the Gibbs sampler as our conditionally optimal Kalman mixture approximation, they derive draws from the predictive density by using a simulation method suggested in Cogley et al. (2005, Section 2.1). More specifically, for each posterior draw $m = 1, \dots, M$, first they generate $\tilde{\beta}_{j,t+1}^{(m)}$ by drawing from the normal distribution $\mathcal{N}(\tilde{\beta}_{jt}^{(m)}, 1)$. Based on $\tilde{\beta}_{j,t+1}^{(m)}$, they forecast

$$y_{t+1}^{(m)} = \mathbf{x}_{t+1} \boldsymbol{\beta}^{(m)} + \mathbf{x}_{t+1} \text{Diag}(\sqrt{\theta_1^{(m)}}, \dots, \sqrt{\theta_d^{(m)}}) \tilde{\beta}_{t+1,j}^{(m)} + \mathcal{N}(0, \sigma_{t+1}^{2(m)}),$$

where the same method as above is applied, to forecast $\sigma_{t+1}^{2(m)}$ for the SV specification. Finally, for each time point $t+1$, they perform a nonparametric kernel smoothing algorithm on $y_{t+1}^{(m)}$ for $m = 1, \dots, M$, which is then evaluated at the observed value y^{t+1} . They use the function `ksdensity` in Matlab, which returns a probability density estimate for the sample in the vector $y_{t+1}^{(m)}$, evaluated at the specified value y^{t+1} .

We also experimented with what we call the *naive Gaussian mixture approximation*. For this approximation, $(\sigma_{t+1}^2)^{(m)}$ and $\beta_{t+1}^{(m)}$ are sampled from the predictive posterior $p(\beta_{t+1}, \sigma_{t+1}^2 | \mathbf{y}^t)$ as

in Belmonte et al. (2014), however the exact predictive density $y_{t+1}|\boldsymbol{\beta}_{t+1}, \sigma_{t+1}^2 \sim \mathcal{N}(\mathbf{x}_{t+1}\boldsymbol{\beta}_{t+1}, \sigma_{t+1}^2)$ is used as component density. This yields

$$\begin{aligned} p(y_{t+1}|\mathbf{y}^t) &= \int p(y_{t+1}|\boldsymbol{\beta}_{t+1}, \sigma_{t+1}^2) p(\boldsymbol{\beta}_{t+1}, \sigma_{t+1}^2|\mathbf{y}^t) d(\boldsymbol{\beta}_{t+1}, \sigma_{t+1}^2) \\ &\approx \frac{1}{M} \sum_{m=1}^M f_N\left(y_{t+1}; \mathbf{x}_{t+1}\boldsymbol{\beta}_{t+1}^{(m)}, (\sigma_{t+1}^2)^{(m)}\right). \end{aligned} \quad (25)$$

Although the naive Gaussian mixture approximation (25) avoids kernel density estimation, we found that it can give very imprecise results, in particular, if the error variance $(\sigma_{t+1}^2)^{(m)}$ is considerably smaller than the forecasting variance $S_{t+1}^{(m)}$ in the conditionally exact Kalman filter approximation (24), see Frühwirth-Schnatter (1992) for an earlier discussion of this problem and Subsection 7.1.2 for more details.

5 Extension to multivariate time series

5.1 TVP models for multivariate time series

The methods introduced in the previous sections are easily extended to TVP models for multivariate time series, such as time-varying parameter VARs, see e.g. Eisenstat et al. (2014) who analyze the response of macro variables to fiscal shocks, and time-varying structural VARs, see e.g. Primiceri (2005) for a monetary policy application. Consider, as illustration, the following TVP model for an r -dimensional time series \mathbf{y}_t ,

$$\mathbf{y}_t = \mathbf{B}_t \mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t), \quad (26)$$

where \mathbf{x}_t is a (column) vector of d regressors, and \mathbf{B}_t is a time-varying $(r \times d)$ matrix with coefficient $\beta_{ij,t}$ in row i and column j , potentially containing structural zeros or constant values, i.e. $\beta_{ij,t} \equiv c$ apriori. The (apriori) unconstrained time-varying coefficients $\beta_{ij,t}$ are assumed to follow independent random walks as in the univariate case:

$$\beta_{ij,t} = \beta_{ij,t-1} + \omega_{ij,t}, \quad \omega_{ij,t} \sim \mathcal{N}(0, \theta_{ij}), \quad (27)$$

with initial value $\beta_{ij,0} \sim \mathcal{N}(\beta_{ij}, \theta_{ij} P_{0,ijj})$, where $P_{0,ijj} \sim \mathcal{G}^{-1}(\nu_P, (\nu_P - 1)c_P)$ as before. Both the fixed regression coefficient β_{ij} as well as the process variance θ_{ij} are assumed to be unknown.

Each of the – apriori unconstrained – coefficients $\beta_{ij,t}$ is potentially constant, in which case the corresponding unknown process variance θ_{ij} is equal to 0. A constant coefficient $\beta_{ij,t} \equiv \beta_{ij}$ is potentially insignificant, in which case $\beta_{ij} = 0$. Hence, shrinkage priors as introduced in Subsection 2.3 and 2.4 for the univariate case, can be immediately imposed on θ_{ij} and β_{ij} to identify which of these scenarios holds for each coefficient $\beta_{ij,t}$.

For hierarchical priors, a decision is required as to whether the hyperparameters κ^2 and λ^2 should be the same for all coefficients. However, for computational reasons, it is advantageous to assume row specific hyperparameters κ_i^2 and λ_i^2 , drawn from a gamma hyper prior. Hence, the triple gamma prior for the process variances θ_{ij} of the coefficients in the i th row of a multivariate TVP model reads:

$$\theta_{ij}|\xi_{ij}^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_{ij}^2}\right), \quad \xi_{ij}^2|a^\xi, \kappa_i^2 \sim \mathcal{G}\left(a^\xi, a^\xi \kappa_i^2/2\right), \quad \kappa_i^2 \sim \mathcal{G}(d_1, d_2),$$

with an individual prior expectation ξ_{ij}^2 for each process variance θ_{ij} . Similarly, an individual prior variance τ_{ij}^2 is introduced for each fixed regression coefficient β_{ij} in the spirit of (15):

$$\beta_{ij}|\tau_{ij}^2 \sim \mathcal{N}(0, \tau_{ij}^2), \quad \tau_{ij}^2|a^\tau, \lambda_i^2 \sim \mathcal{G}(a^\tau, a^\tau \lambda_i^2/2), \quad \lambda_i^2 \sim \mathcal{G}(e_1, e_2).$$

If the errors ε_t are uncorrelated, i.e. Σ_t is a diagonal matrix, then the multivariate TVP model (26) has a representation as r independent univariate TVP parameters models as in Subsection 2.1 and MCMC estimation as introduced in Section 3 can be performed independently for all r rows of the system. If Σ_t is a full covariance matrix, then the rows are not independent, because of the correlation among the various components in ε_t . However, as shown by Lopes et al. (2015), a Cholesky decomposition of Σ_t leads to such a representation, see also Eisenstat et al. (2014) and Zhao et al. (2016). Further details are provided in the next subsection.

5.2 The time-varying parameter Cholesky SV model

Lopes et al. (2015) demonstrate how multivariate time series with time-varying covariance matrix, following

$$\mathbf{y}_t \sim \mathcal{N}_r(\mathbf{0}, \Sigma_t), \tag{28}$$

can be transformed into a system of r independent equations, using the time-varying Cholesky decomposition $\Sigma_t = \mathbf{A}_t \mathbf{D}_t \mathbf{A}_t'$, where $\mathbf{A}_t \mathbf{D}_t^{1/2}$ is the lower triangular Cholesky decomposition of Σ_t . Here, \mathbf{A}_t is also lower triangular with ones in the main diagonal, while \mathbf{D}_t is a time-varying diagonal matrix. It follows that

$$\mathbf{A}_t^{-1} \mathbf{y}_t \sim \mathcal{N}_r(\mathbf{0}, \mathbf{D}_t). \tag{29}$$

Setting the elements of \mathbf{A}_t^{-1} equal to $\Phi_{ij,t}$, for $j < i$, it can be shown that equation (29) can be

decomposed into

$$\begin{pmatrix} 1 & \dots & & & 0 \\ \Phi_{21,t} & 1 & & & 0 \\ & & \ddots & & 0 \\ \vdots & & & 1 & 0 \\ \Phi_{r1,t} & \Phi_{r2,t} & \dots & \Phi_{r,r-1,t} & 1 \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{rt} \end{pmatrix} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{D}_t),$$

which can be written as in (26):

$$\mathbf{y}_t \sim \mathcal{N}_r(\mathbf{B}_t \mathbf{x}_t, \mathbf{D}_t), \quad (30)$$

where \mathbf{B}_t is a $r \times (r-1)$ matrix with elements $\beta_{ij,t} = -\Phi_{ij,t}$, \mathbf{D}_t is a diagonal matrix and the $(r-1)$ -dimensional vector $\mathbf{x}_t = (y_{1t}, \dots, y_{r-1,t})'$ is a regressor derived from \mathbf{y}_t . Thus the joint normal distribution of \mathbf{y}_t can be written as a system of r independent TVP models as introduced in Subsection 5.1, where each time-varying coefficients $\beta_{ij,t}, j < i, i = 1, \dots, r$, follows a random walk as in (27).

To capture conditional heteroscedasticity, the diagonal matrix $\mathbf{D}_t = \text{Diag}(e^{h_{1t}}, \dots, e^{h_{rt}})$ is assumed to be time-varying, where for each row $i = 1, \dots, r$, the log volatility h_{it} is assumed to follow an individual SV model as in (5), with row specific parameters μ_i, ϕ_i , and $\sigma_{\eta,i}^2$:

$$h_{it}|h_{i,t-1}, \mu_i, \phi_i, \sigma_{\eta,i}^2 \sim \mathcal{N}(\mu_i + \phi_i(h_{i,t-1} - \mu_i), \sigma_{\eta,i}^2).$$

For $r = 3$, for instance, the TVP Cholesky SV model reads:

$$\begin{aligned} y_{1t} &= \varepsilon_{1t}, & \varepsilon_{1t} &\sim \mathcal{N}(0, e^{h_{1t}}), \\ y_{2t} &= \beta_{21,t}y_{1t} + \varepsilon_{2t}, & \varepsilon_{2t} &\sim \mathcal{N}(0, e^{h_{2t}}), \\ y_{3t} &= \beta_{31,t}y_{1t} + \beta_{32,t}y_{2t} + \varepsilon_{3t}, & \varepsilon_{3t} &\sim \mathcal{N}(0, e^{h_{3t}}). \end{aligned}$$

Note that for all TVP models no intercept is present. For the first TVP model, no regressors are present and only the time-varying volatility h_{1t} has to be estimated. In the i -th equation, $i-1$ regressors are present and $d = i-1$ time-varying regression coefficients $\beta_{ij,t}$ as well as the time-varying volatilities h_{it} needs to be estimated. Each of these equations can be transformed in an obvious manner into a non-centered state space model and the efficient MCMC scheme developed in Algorithm 1 can be applied to perform Bayesian inference under the hierarchical shrinkage priors on β_{ij} and θ_{ij} .

6 Illustrative Application to Simulated Data

In this section, we illustrative our proposed methodology for simulated data. We generate 100 univariate time series of length $T = 200$ from the state space model introduced in Subsection 2.1 where $d = 3$ and the regressors $\{x_{2t}\}$ and $\{x_{3t}\}$ are iid draws from a standard normal distribution. We set the true values $(\beta_1, \beta_2, \beta_3) = (-3.5, 0.4, 0.004)$ and $(\sqrt{\theta_1}, \sqrt{\theta_2}, \sqrt{\theta_3}) = (0.0062, 0.05, 0.0001)$, which yields one constant but significant state (β_{1t}), one heavily time-varying latent state (β_{2t}), and one insignificant state (β_{3t}). We assume a homoscedastic error model with $\sigma^2 = 1$.

We consider shrinkage priors on $\sqrt{\theta_j}$ and β_j and compare $a^\tau = a^\xi = 0.1$ with $a^\tau = a^\xi = 1$ (which corresponds to the Bayesian Lasso) under the hyperparameter setting $d_1 = d_2 = e_1 = e_2 = 1$. For each of the 100 simulated time series, we estimate the TVP model under both shrinkage priors using Algorithm 1, by drawing $M = 30,000$ samples after a burn-in of length 30,000.

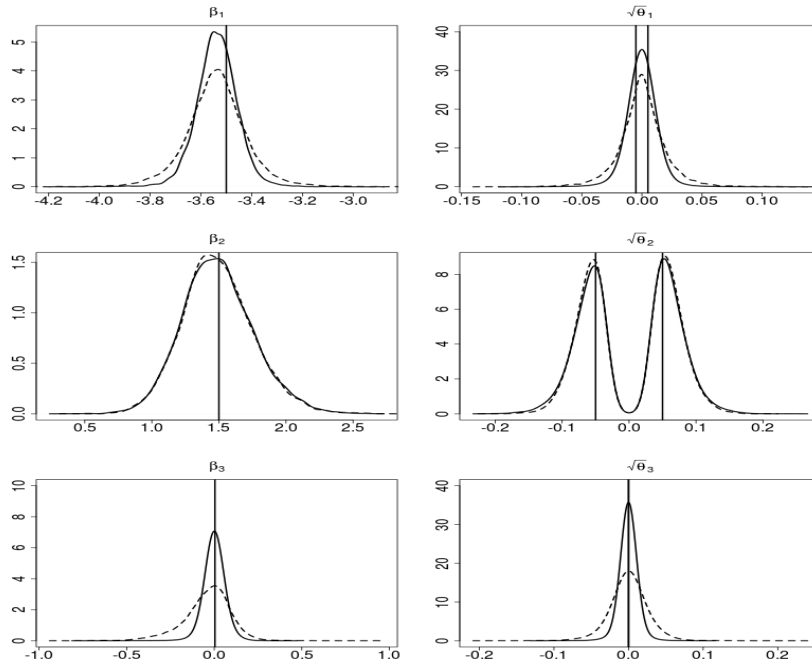


Figure 4: Simulated data. Posterior densities of β_j (left hand side) and $\sqrt{\theta_j}$ (right hand side) together with the true values (indicated by the vertical lines), based on shrinkage priors with $a^\tau = a^\xi = 0.1$ (solid line) and $a^\tau = a^\xi = 1$ (dashed line).

In Figure 4 we present the posterior densities for β_j and $\sqrt{\theta_j}$ for one such time series for the hyperparameter setting $a^\tau = a^\xi = 0.1$ (solid line) and compare it to the hyperparameter setting $a^\tau = a^\xi = 1$ (dashed line), together with the true values. For the coefficients with the small variances $\sqrt{\theta_1}$ and $\sqrt{\theta_3}$, the posterior density is shrunk towards zero, while the bimodal structure of the posterior density is well pronounced for the relatively larger value $\sqrt{\theta_2}$. Further,

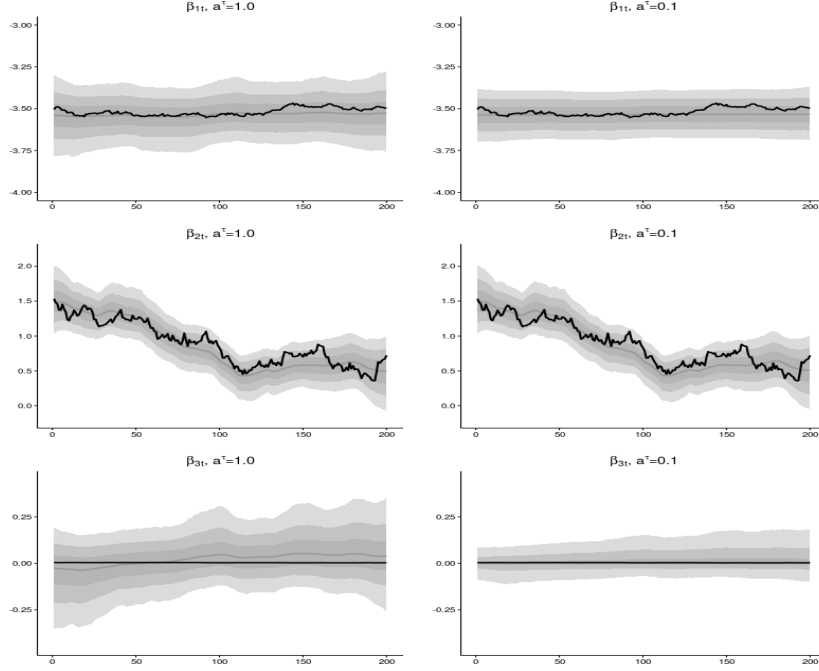


Figure 5: Simulated data. Pointwise (0.025, 0.25, 0.5, 0.75, 0.975)-quantiles of the posterior paths $\beta_{jt} = \beta_j + \sqrt{\theta_j} \tilde{\beta}_{jt}$ in the centered parametrization in comparison to the true paths (thick black line) for one of the simulated time series; left hand side: $a^\tau = a^\xi = 1$, right hand side: $a^\tau = a^\xi = 0.1$.

we show the posterior paths of β_{jt} in Figure 5. Evidently, our algorithm is able to detect the heavily time-varying latent state β_{2t} , the constant but significant state β_{1t} and the insignificant state β_{3t} . In both figures, the advantage of using the normal-gamma prior as opposed to the Bayesian Lasso is reflected by increased efficiency in identifying coefficients that are not time-varying.

As a summary, we present for all parameters $\beta_1, \beta_2, \beta_3$ and $|\sqrt{\theta_1}|, |\sqrt{\theta_2}|, |\sqrt{\theta_3}|$ the average mean squared error ($avMSE$), the average squared bias ($avBIAS^2$) and the average variance ($avVAR$) over the 100 simulated time series in Table 1.⁶ Clearly, for the two coefficients which are not fully time-varying, heavier shrinkage introduced by a prior with $a^\tau = a^\xi = 0.1$ leads to reduced $avMSE$ compared to $a^\tau = a^\xi = 1$.

⁶Given M draws $\vartheta^{(i1)}, \dots, \vartheta^{(iM)}$ of a parameter ϑ for each time series i , these measures are defined as:

$$avMSE = avVAR + avBIAS^2, \quad avVAR = \frac{1}{100} \sum_{i=1}^{100} V_i, \quad avBIAS^2 = \frac{1}{100} \sum_{i=1}^{100} (E_i - \vartheta^{\text{true}})^2, \\ E_i = \frac{1}{M} \sum_{m=1}^M \vartheta^{(im)}, \quad V_i = \frac{1}{M} \sum_{m=1}^M \left(\vartheta^{(im)} - E_i \right)^2.$$

	$a^\tau = a^\xi = 0.1$			$a^\tau = a^\xi = 1$		
Name	$avMSE$	$avVAR$	$avBIAS^2$	$avMSE$	$avVAR$	$avBIAS^2$
β_1	7.7E-03	6.9E-03	7.7E-04	1.6E-02	1.5E-02	9.4E-04
β_2	1.8E-01	6.0E-02	1.2E-01	1.5E-01	5.5E-02	9.3E-02
β_3	2.9E-03	2.5E-03	4.0 E-04	2.7E-02	2.0E-02	6.5E-03
$ \sqrt{\theta_1} $	6.9E-05	6.8E-05	1.7E-06	2.7E-04	1.9E-04	8.3E-05
$ \sqrt{\theta_2} $	9.0E-04	6.2E-04	2.8E-04	6.4E-04	4.9E-04	1.5E-04
$ \sqrt{\theta_3} $	1.5 E-04	1.1 E-04	3.8E-05	7.0E-04	3.0E-04	4.0E-04

Table 1: Simulated data. Average mean squared error ($avMSE$), average variance ($avVAR$), and average squared bias ($avBIAS^2$) over 100 simulated time series.

7 Applications in Economics and Finance

With our real world applications, we were confronted with a number of implementation issues which led us to include the interweaving step as well as exploring different forecasting schemes.

7.1 Inflation modelling

As a first application, we reconsider EU-area inflation data analysed in Belmonte et al. (2014) and consider the generalized Phillips curve specification, where inflation π_t depends on (typically $p = 12$) lags of inflation and other predictors x_t :

$$\pi_{t+h} = \sum_{j=0}^{p-1} \phi_{jt} \pi_{t-j} + \gamma_t x_t + \varepsilon_{t+h}, \quad \varepsilon_{t+h} \sim \mathcal{N}(0, \sigma^2). \quad (31)$$

This set up has been discussed by Stock and Watson (2012), among others, for forecasting both the annual inflation rate ($h = 12$) and the monthly inflation rate ($h = 1$). Data are monthly and range from February 1994 until November 2010. We list precise definitions of these variables in Appendix A.2.1. As we work with time series which are not seasonally adjusted we include time-varying dummy variables in the right hand side of (31) to account for seasonal patterns. Thus we are estimating in total $d = 37$ possibly time-varying coefficients, consisting of the intercept, 13 regressors like the *unemployment rate* and the *1-month interest rate*, 12 lagged values of inflation and eleven seasonal dummies.

We investigated shrinkage priors with shrinkage parameter $a^\tau = a^\xi$ in the range 0.1 to 1. Concerning the hyperparameter settings, we studied a wide range of combinations from $d_1 = d_2 = e_1 = e_2 = 0.001$ to $d_1 = d_2 = e_1 = e_2 = 1$. The setting where $a^\tau = a^\xi = 1$ and $d_1 = d_2 = e_1 = e_2 = 0.001$ corresponds to the choice recommended by Belmonte et al. (2014). For each of these settings, MCMC inference is based on Algorithm 1 including interweaving, with $M = 100,000$ draws after a burn-in of the same size.

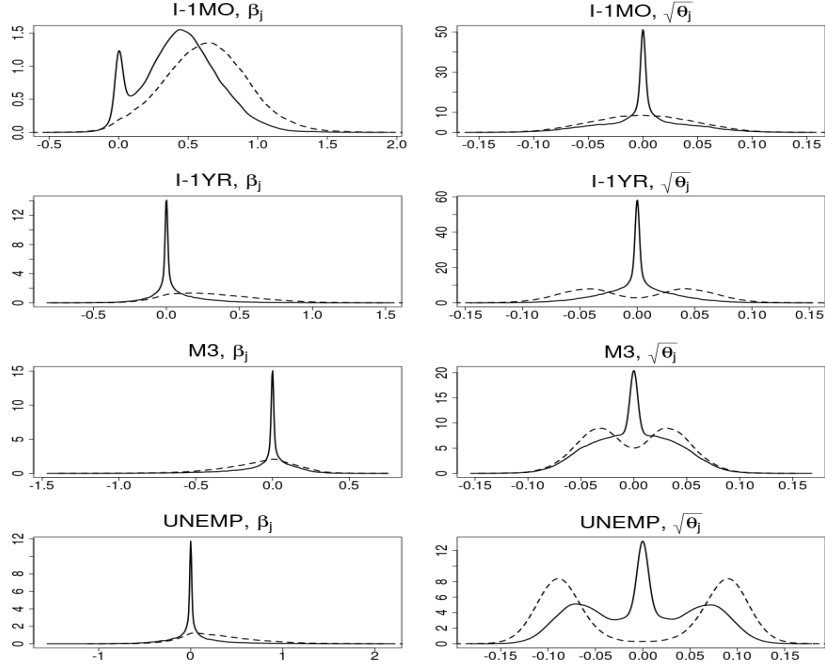


Figure 6: ECB data. Posterior densities of β_j (left hand side) and $\sqrt{\theta_j}$ (right hand side), based on shrinkage priors with $a^\tau = a^\xi = 0.2$ (solid line) and $a^\tau = a^\xi = 1$ (dashed line) and $d_1 = d_2 = e_1 = e_2 = 1$ for following predictors (from top to bottom): *1-month interest rate*, *1-year interest rate*, *M3*, and *unemployment rate*.

7.1.1 Estimation results

In general, we want to distinguish three types of parameters β_{jt} : (a) time-varying, (b) static but significant, and (c) insignificant. One way to achieve such a classification is simply by visual inspection of the posterior distribution of β_j and $\sqrt{\theta_j}$. The posterior density of any scale parameter $\sqrt{\theta_j}$ needs to be symmetric around zero. Thus, if the unknown variance θ_j is systematically different from zero, then the posterior density of $\sqrt{\theta_j}$ is likely to be bimodal. If we find that the posterior density of $\sqrt{\theta_j}$ is unimodal, then the unknown variance is likely to be zero.

The easiest combination to spot is if both parameters β_j and $\sqrt{\theta_j}$ are shrunk towards zero and the corresponding posterior densities exhibit peaks at zero. This is the case for most of the 37 parameters. Those exhibiting a more interesting behaviour are the *1-month interest rate* ($j = 14$), the *1-year interest rate* ($j = 15$), *M3* ($j = 22$), and the *unemployment rate* ($j = 26$). For illustration, we present the corresponding posterior density estimates of β_j and $\sqrt{\theta_j}$ in Figure 6 for the hyperparameter setting $a^\tau = a^\xi = 0.2$ and $a^\tau = a^\xi = 1$ with $d_1 = d_2 = e_1 = e_2 = 1$. As expected, we find a bimodal structure in some cases and the corresponding posterior paths $\beta_{jt} = \beta_j + \sqrt{\theta_j} \tilde{\beta}_{jt}$, reported in Figure 7 for $a^\tau = a^\xi = 0.2$, show that these parameters are, indeed, time-varying.

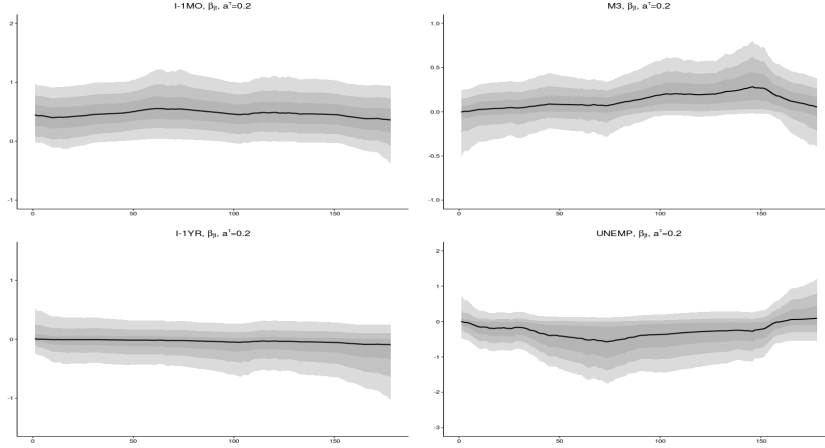


Figure 7: ECB data. $(0.025, 0.25, 0.5, 0.75, 0.975)$ -quantiles of the posterior paths of $\beta_{jt} = \beta_j + \sqrt{\theta_j} \tilde{\beta}_{jt}$, based on the shrinkage prior with $a^\tau = a^\xi = 0.2$ and $d_1 = d_2 = e_1 = e_2 = 1$; left hand side: *1-month interest rate* (top) and *1-year interest rate* (bottom); right hand side: *M3* (top) and *unemployment rate* (bottom).

For $a^\tau = a^\xi = 0.2$, the path of the *1-month interest rate* is significantly different from zero. There is a mode close to zero for β_j , the mean and the median of β_j is 0.44 and the posterior density of $\sqrt{\theta_j}$ exhibits a peak at zero. The *1-year interest rate* is basically shrunk towards zero and can be regarded as insignificant. A time-varying behaviour is visible for *M3* and the *unemployment rate*. Summary statistics are given in Table A.2 in Appendix A.3.

$a^\tau = a^\xi = 0.2$					$a^\tau = a^\xi = 1$			
no ASIS		ASIS			no ASIS		ASIS	
j	β_j	$ \sqrt{\theta_j} $	β_j	$ \sqrt{\theta_j} $	β_j	$ \sqrt{\theta_j} $	β_j	$ \sqrt{\theta_j} $
1	3641	211	133	94	1749	158	97	67
14	248	271	82	149	84	110	38	83
15	101	240	70	134	110	135	55	76
22	172	220	64	140	119	96	45	63
26	630	412	89	238	485	158	63	55

Table 2: ECB data. Inefficiency factors of MCMC posterior draws of selected parameters, obtained from Algorithm 1 with and without interweaving under shrinkage priors with $d_1 = d_2 = e_1 = e_2 = 1$ and, respectively, $a^\tau = a^\xi = 0.2$ and $a^\tau = a^\xi = 1$.

For this data set, full conditional MCMC sampling turned out to be extremely inefficient and basically resulted in a failure to converge in a reasonable amount of time. This problem motivated us to include the interweaving step in the Gibbs sampler outlined in Algorithm 1. As illustrated for selected parameters in Table 2, adding the interweaving step leads to substantial improvement of the mixing behaviour of MCMC sampling, with considerably reduced inefficiency factors. For further illustration, we show MCMC paths obtained for β_1 with and without interweaving in

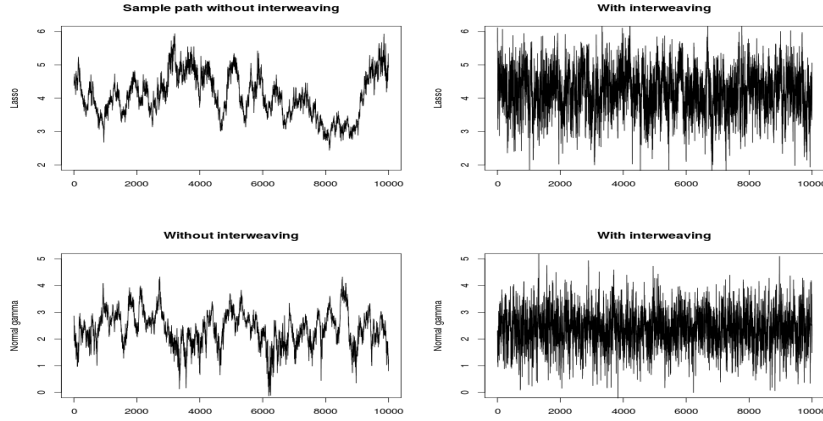


Figure 8: ECB data. Sample paths of β_1 comparing the MCMC schemes without interweaving (left hand side) and with interweaving (right hand side) for $a^\xi = a^\tau = 1$ (top row) and $a^\xi = a^\tau = 0.2$ (bottom row). $M = 100,000$ draws, only every tenth draw is shown.

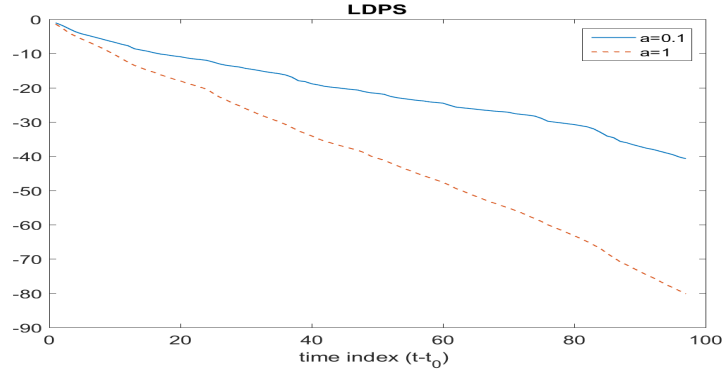


Figure 9: ECB data. Cumulative log predictive scores for the last 100 points in time obtained through the conditionally optimal Kalman mixture approximation under $a^\tau = a^\xi = 0.1$ (solid line) and $a^\tau = a^\xi = 1$ (dashed line) with $d_1 = d_2 = e_1 = e_2 = 0.001$.

Figure 8.

7.1.2 Predictive performance

Here we evaluate the predictive performance of our model and show that by incorporating the normal-gamma prior as a shrinkage prior we improve the predictive performance behaviour. Figure 9 shows the cumulative log predictive scores over the last 100 time points, using the conditionally optimal Kalman mixture approximation derived in Subsection 4.2. Evidently, for this time series, the predictive performance can be improved considerably by using the normal-gamma prior as a shrinkage prior rather than the hierarchical Bayesian Lasso prior applied by Belmonte et al. (2014).

For numerical reasons, it is essential to use an accurate approximation method such as the conditionally optimal Kalman mixture approximation rather than the naive approximation discussed in Subsection 4.2 to approximate the predictive density. Since both approximations are finite mixtures of Gaussian distributions, the value of each mixture component pdf needs to be calculated at the actually observed value y_{t+1} in order to derive the log predictive scores. Despite the high number of mixture components ($M = 100,000$), the resulting approximations of LPDS_{t+1}^* turned out to differ substantially, see Figure A.2 in Appendix A.3, where a comparison is provided for $a^\tau = a^\xi = 0.1$. In Figure 10 we compare (a small subset of) the mixture component pdfs for both approximations for a single point in time. Compared to the conditionally optimal Kalman mixture approximation, the mixture components of the naive approximation have a very small spread and a large number of zeros results from density evaluation. This leads to a bias of the corresponding estimate of LPDS_{t+1}^* toward 0 and, depending on the true value, the log predictive density score LPDS_{t+1}^* is over- or underrated.

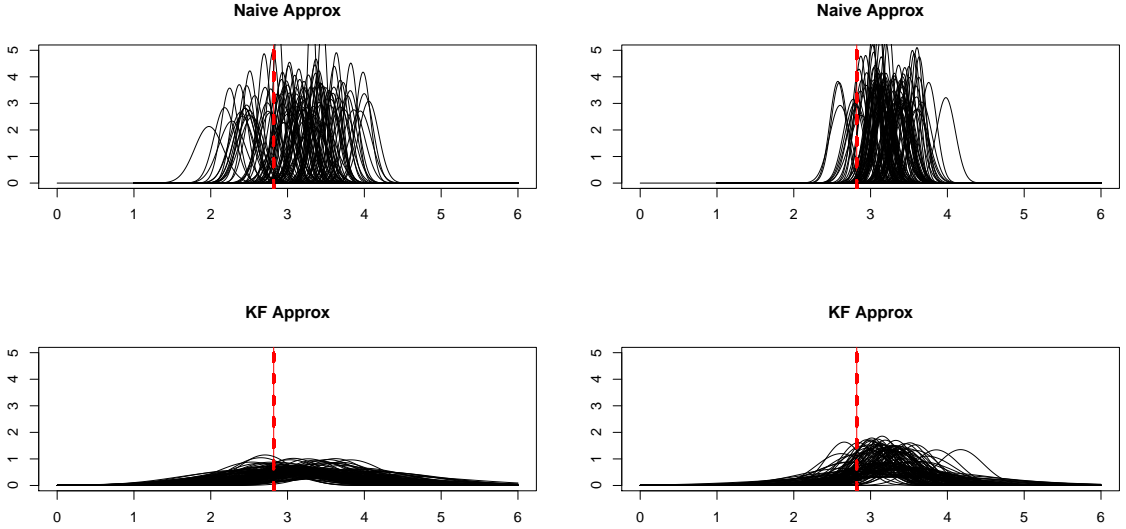


Figure 10: ECB data. Illustration of the naive mixture approximation and the conditionally optimal Kalman mixture approximation for the predictive density at $t = t_0 + 40$ for $a^\tau = a^\xi = 0.1$ (left hand side) and $a^\tau = a^\xi = 1$ (right hand side) for $d_1 = d_2 = e_1 = e_2 = 0.001$. For each, only 40 predictive densities are drawn for better readability. The actually observed value y_{t+1} is indicated by a vertical line.

7.2 TVP Cholesky SV modelling of DAX returns

As a second real world data application, we fit the time-varying parameter Cholesky SV model introduced in Subsection 5.2 to 29 indices from the German Stock Index DAX. The ordering of the indices is alphabetically and our data set spans roughly 2500 daily stock returns from

September 4th, 2001 until August 31st, 2011. More details are available in Appendix A.2.2.

As outlined in Subsection 5.2, we estimate a pure stochastic volatility model for the first index and 28 time-varying parameter regression models with SV error specification for the remaining indices, where the dimension d is increasing from 1 to 28. To estimate the resulting 406, potentially time-varying, coefficients $\beta_{ij,t}$ in an efficient manner, we apply the shrinkage priors introduced in this paper. We investigated a wide range of different hyperparameter settings $a^\tau = a^\xi$, including 0.05, 0.1, \dots , 1, for $d_1 = d_2 = e_1 = e_2 = 0.001$. For all priors, MCMC inference is performed using Algorithm 1 with $M = 50,000$ draws after a burn-in of 50,000.

In addition to our shrinkage priors, we apply the usual conditionally conjugate prior, i.e. $\theta_{ij} \sim \mathcal{IG}(s_0, S_0)$ for all process variances θ_{ij} and $\beta_{ij} \sim \mathcal{N}(0, A_0)$ for all fixed regression coefficients β_{ij} , with prior setting as in Petris et al. (2009), namely $s_0 = 0.1$, $S_0 = 0.001$ and $A_0 = 10$.⁷

Exemplarily detailed results are presented for the tenth time-varying regression in Figure 11, where we compare the posterior densities of β_{ij} and $\sqrt{\theta_{ij}}$, for $i = 10$ and $j = 1, \dots, 9$, obtained under shrinkage priors with $a^\tau = a^\xi = 1$ and $a^\tau = a^\xi = 0.1$ and the conditionally conjugate prior. As expected, under the conditionally conjugate prior all posteriors distributions of $\sqrt{\theta_{ij}}$ are bounded away from 0, with the shape and the position of the symmetric posterior modes (roughly ± 0.015) being more or less the same for all coefficients. Location and shape of the posterior distribution is mainly driven by the prior – for the alternative hyperparameters $s_0 = 0.5$ and $S_0 = 0.2275$ (not shown in the figure) the posterior modes shift to around ± 0.1 .

As opposed to this, the shrinkage priors introduced in this paper allow the posterior distribution of $\sqrt{\theta_{ij}}$ to concentrate at 0, if appropriate. Investigating the posterior distributions of $\sqrt{\theta_{ij}}$ under these shrinkage priors clearly allows to distinguish between components that are really time-varying ($j = 1, 2, 7$) and the remaining components which turn out to be static, both for $a^\tau = a^\xi = 0.1$ and $a^\tau = a^\xi = 1$. For the static components, the posterior distributions of β_{ij} indicate that some coefficients are clearly significant, in particular $j = 3$ and $j = 9$, whereas others are clearly insignificant, in particular $j = 6$ and $j = 8$.

These findings are confirmed by the corresponding posterior paths of $\beta_{ij,t} = \beta_{ij} + \sqrt{\theta_{ij}}\tilde{\beta}_{ij,t}$ which are displayed for all components $j = 1, \dots, 9$ in Figure 12 under the shrinkage prior (with $a^\tau = a^\xi = 0.1$) and the $\mathcal{IG}(0.1, 0.001)$ -prior. The coefficients $\beta_{i1,t}$, $\beta_{i2,t}$, and $\beta_{i7,t}$ are the only ones that are clearly time-varying, whereas $\beta_{i3,t}$ and $\beta_{i9,t}$ are constant, but clearly shifted away from 0.

⁷MCMC estimation under this prior requires a minor modification of Algorithm 1. We sample θ_{ij} in the centered parameterization, only, from $\theta_{ij} \sim \mathcal{IG}\left(s_0 + \frac{T+1}{2}, S_0 + \frac{1}{2} \sum_{t=1}^T (\beta_{ij,t} - \beta_{ij,t-1})^2 + \frac{(\beta_{ij,0} - \beta_{ij})^2}{2P_{0,ijj}}\right)$.

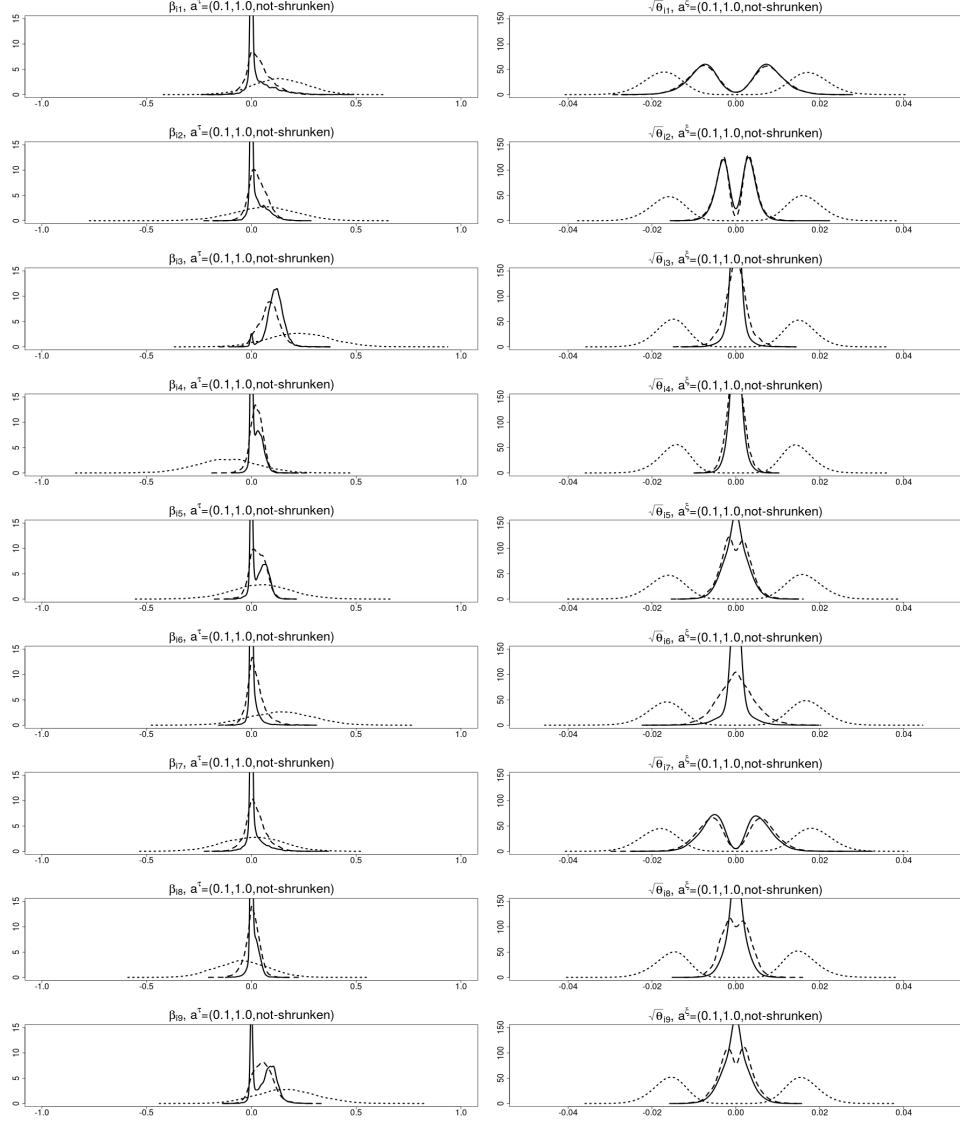


Figure 11: DAX data. Posterior densities of β_{ij} (left hand side) and $\sqrt{\theta_{ij}}$ (right hand side) for $i = 10$ and $j = 1, \dots, 9$ (from top to bottom), derived under the conditionally conjugate prior $\beta_{ij} \sim \mathcal{N}(0, 10)$ and $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (dotted line) and shrinkage priors with $a^\tau = a^\xi = 0.1$ (solid line) and $a^\tau = a^\xi = 1$ (dashed line) and $d_1 = d_2 = e_1 = e_2 = 0.001$.

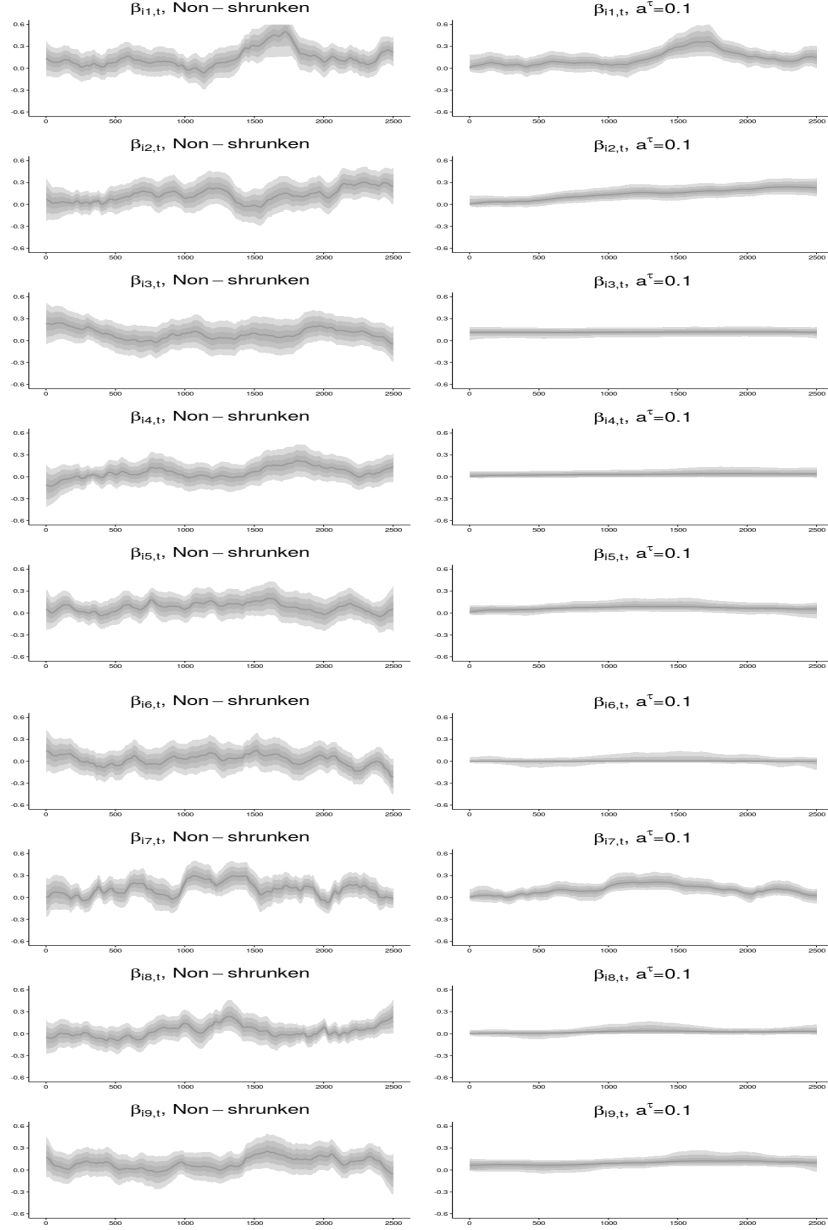


Figure 12: DAX data. (0.025, 0.25, 0.5, 0.75, 0.975)-quantiles of the posterior paths $\beta_{ij,t} = \beta_{ij} + \sqrt{\theta_{ij}}\tilde{\beta}_{ij,t}$ in the tenth TVP regression (i.e. $i = 10$) and $j = 1, \dots, 9$ (from top to bottom); derived under the conditionally conjugate prior $\beta_{ij} \sim \mathcal{N}(0, 10)$ and $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (left hand side) and a shrinkage prior with $a^\tau = a^\xi = 0.1$ and $d_1 = d_2 = e_1 = e_2 = 0.001$ (right hand side).

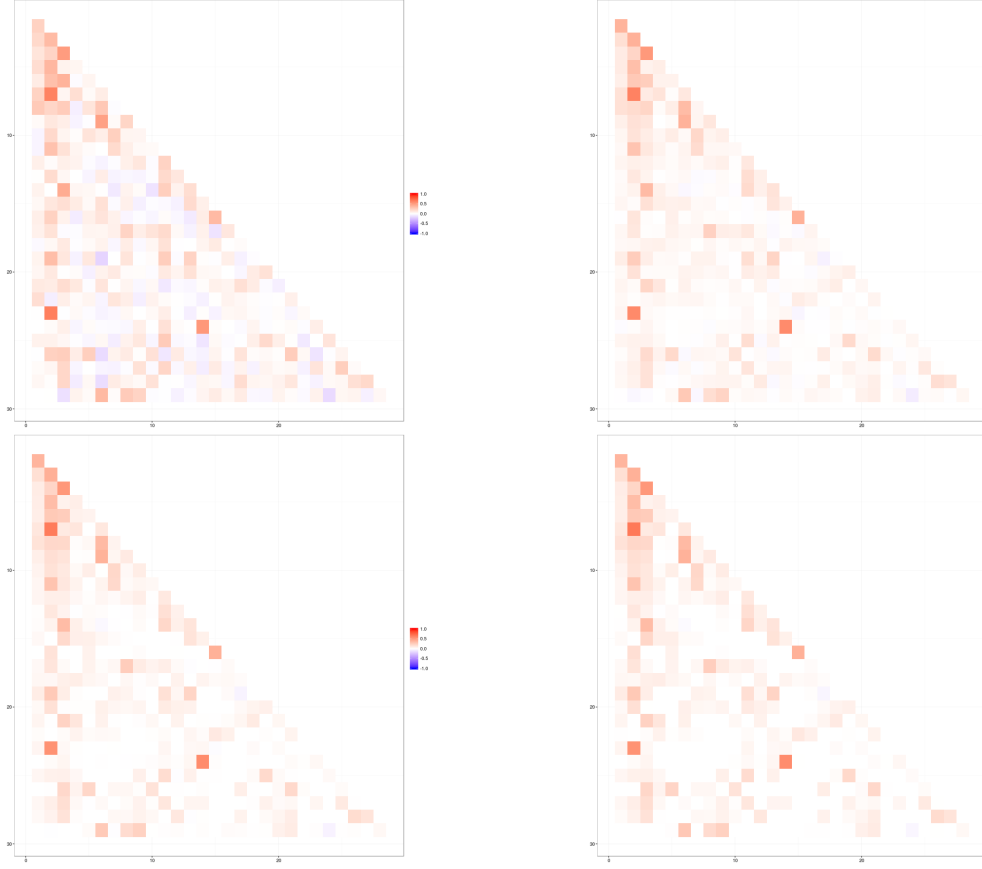


Figure 13: DAX data. Heat plot of the posterior median of the 29×28 Cholesky factor matrix \mathbf{B}_t at $t = 1150$, derived under the conditionally conjugate prior $\beta_{ij} \sim \mathcal{N}(0, 10)$ and $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (top left) and shrinkage priors with $d_1 = d_2 = e_1 = e_2 = 0.001$ and $a^\tau = a^\xi = 1$ (top right), $a^\tau = a^\xi = 0.1$ (bottom left) and $a^\tau = a^\xi = 0.05$ (bottom right). Red and blue indicates, respectively, positive and negative values. Values close to zero are white.

Figure 12 also demonstrates a dramatic gain in statistical efficiency, in terms of dispersion of the posterior distribution of $\beta_{ij,t}$ for each point in time, compared to the conditionally conjugate prior. This holds in particular for coefficients which are static, but significant such as $\beta_{i3,t}$ and $\beta_{i9,t}$. In addition to this efficiency again, the estimated paths are much smoother under the shrinkage prior, which facilitates the interpretation of the time-varying components $\beta_{i1,t}$, $\beta_{i2,t}$, and $\beta_{i7,t}$. The coefficient $\beta_{i2,t}$, for instance, shows a trending behaviour, which is not apparent under the conditionally conjugate prior.

When comparing both shrinkage priors in Figure 11, the influence of the increased shrinkage introduced by $a^\tau = a^\xi = 0.1$ is evident. For static components, the posterior of $\sqrt{\theta}_{ij}$ under $a^\tau = a^\xi = 0.1$ shows a much more pronounced spike at 0 for than for the Lasso prior $a^\tau = a^\xi = 1$, which leads to increased efficiency in identifying static components.

Similar impact of our shrinkage method can be observed for the remaining 27 equations

in the TVP model. Overall, we investigated all 406 posterior paths $\beta_{ij,t}$, together with the corresponding posterior distributions of $\sqrt{\theta}_{ij}$ and β_{ij} , and found that a large fraction of these coefficients is not significant. For illustration, we display in Figure 13 one (out of 2500) heat plot of the posterior median of the 29×28 Cholesky factor matrix \mathbf{B}_t at $t = 1150$ for various priors. Whereas the majority of the estimated coefficients $\hat{\beta}_{ij,t}$ is different from 0 for the conditionally conjugate prior, only a small part is significantly different from 0 for the various shrinkage priors, based on $a^\tau = a^\xi = 0.05$, $a^\tau = a^\xi = 0.1$, and $a^\tau = a^\xi = 1$.

The Cholesky factor matrices \mathbf{B}_t displayed in Figure 13 indicate little difference between the various shrinkage priors. Detailed numerical results, shown in Appendix A.3 in Figure A.4 reveal that all shrinkage priors identify more or the less the same coefficients $\hat{\beta}_{ij,t}$ as being insignificant. The various priors, however, have an impact on all non-zero coefficients and, as a consequence, on the off-diagonal elements and other functionals of the resulting time-varying covariance matrix Σ_t .

Finally, we compare various priors using log predictive scores for the last 500 returns, with the first 400 observation serving as training sample. Very conveniently, the triangular structure of the model allows to decompose the 28-dimensional predictive density as $p(\mathbf{y}_{t+1}|\mathbf{y}^t) = \prod_{i=1}^r p(y_{i,t+1}|\mathbf{y}^t)$. Hence, the overall log predictive density score LPDS_{t+1}^* results as the sum of the individual log predictive density scores $\text{LPDS}_{i,t+1}^* = \log p(y_{i,t+1}|\mathbf{y}^t)$, derived independently for each of the $r = 29$ TVP models:

$$\text{LPDS}_{t+1}^* = \log p(\mathbf{y}_{t+1}|\mathbf{y}^t) = \sum_{i=1}^r \log p(y_{i,t+1}|\mathbf{y}^t) = \sum_{i=1}^r \text{LPDS}_{i,t+1}^*.$$

The individual log predictive density scores $\text{LPDS}_{i,t+1}^*$ are approximated using the conditionally optimal Kalman mixture approximation introduced in Subsection 4.2. Interestingly, for this study, the naive Gaussian mixture approximation yields identical estimated of $\text{LPDS}_{i,t+1}^*$ for nearly all time point, see Figure A.3 in Appendix A.3, mainly because the variance of all conditional mixture densities $p(y_{i,t+1}|\mathbf{y}^t, \boldsymbol{\theta})$ is dominated by $\sigma_{i,t+1}^2$ for both approximations.

The cumulative log predictive scores, resulting from the conditionally optimal Kalman mixture approximation, are shown in Figure 14 for various priors. We find overwhelming evidence in favour of using shrinkage priors instead of the conditionally conjugate prior. For the later, the choice of the hyperparameters, e.g. $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ versus $\theta_{ij} \sim \mathcal{IG}(0.5, 0.2275)$, exercises tremendous influence on the log predictive scores, whereas the log predictive scores are much more similar for the various values of $a^\tau = a^\xi$.

8 Conclusion

In the present paper, shrinkage for time-varying parameter (TVP) models was investigated within a Bayesian framework, with the aim to automatically reduce time-varying parameters to

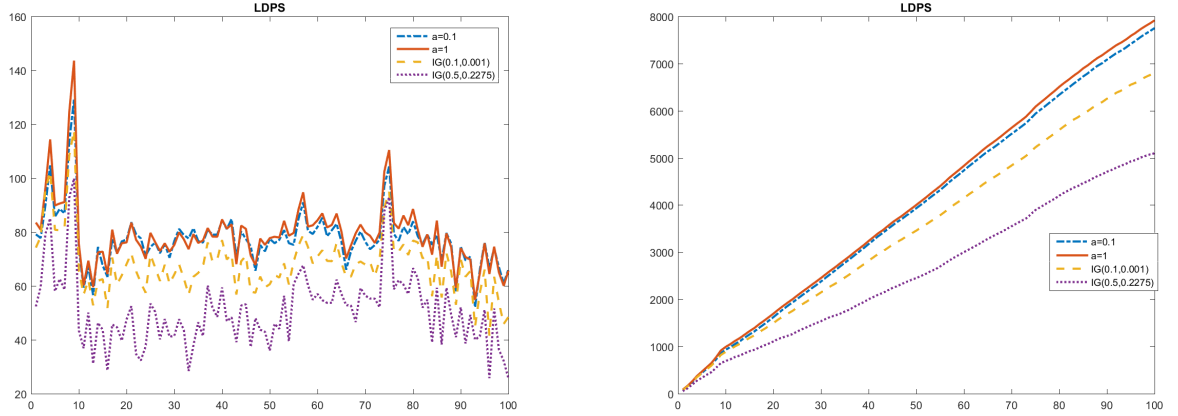


Figure 14: DAX data. Individual (left hand side) and cumulative (right hand side) log predictive density scores for the last 100 time points using the last 400 observations as training sample. Shrinkage prior with $a^\tau = a^\xi = 1$ (full line) and $a^\tau = a^\xi = 0.1$ (dash-dotted line) in comparison to the conditionally conjugate priors $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (dashed line) and $\theta_{ij} \sim \mathcal{IG}(0.5, 0.2275)$ (dotted line).

static ones, if the model is overfitting. This goal was achieved by formulating shrinkage priors based on the normal-gamma prior (Griffin and Brown, 2010), in particular, for the process variances, extending previous work based on spike-and-slab priors (Frühwirth-Schnatter and Wagner, 2010) and the Bayesian Lasso prior (Belmonte et al., 2014).

An efficient MCMC estimation scheme, exploiting boosting ideas such as ancillarity-sufficiency interweaving (Yu and Meng, 2011), was developed and applied to TVP models both for univariate and multivariate time series. Our applications in economics and finance included EU area inflation modelling based on the generalized Phillips curve and estimating a time-varying covariance matrix based on a TVP Cholesky SV model for a multivariate time series of returns derived from the DAX-30 index. We investigated different prior settings, including the common inverted gamma prior for the process variances, using a predictive analysis. Overall, our findings suggest that the family of shrinkage priors introduced in this paper for TVP models is successful in avoiding overfitting, if potentially time-varying parameters are, indeed, static or even insignificant.

References

- Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Ser. B*, 36:99–102.
- Belmonte, M. A. G., Koop, G., and Korobolis, D. (2014). Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33:80–94.

- Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. pages 88–95.
- Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81:541–553.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480.
- Cogley, T., Morozov, S., and Sargent, T. (2005). Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system. *Journal of Economic Dynamics and Control*, 16:1893–1925.
- Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106:157–181.
- Diebold, F. X., Gunter, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883.
- Diks, C., Panchenko, V., and Dijk, D. v. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163:215–230.
- Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Ser. B*, 62:3–56.
- Eisenstat, E., Chan, J. C., and Strachan, R. W. (2014). Stochastic model specification search for time-varying parameter VARs. *SSRN Electronic Journal* 01/2014; DOI: 10.2139/ssrn.2403560.
- Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20:203–219.
- Frühwirth-Schnatter, S. (1992). Approximate predictive integrals for dynamic generalized linear models. In Fahrmeir, L., Francis, B., Gilchrist, R., and Tutz, G., editors, *Advances in GLIM and Statistical Modelling*, number 78 in Lecture Notes in Statistics, pages 123–151. Springer, New York.
- Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202.
- Frühwirth-Schnatter, S. (1995). Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of the Royal Statistical Society, Ser. B*, 57:237–246.

- Frühwirth-Schnatter, S. (2001). Fully Bayesian analysis of switching Gaussian state space models. *Annals of the Institute of Statistical Mathematics*, 53:31–49.
- Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation. In Harvey, A., Koopman, S. J., and Shephard, N., editors, *State Space and Unobserved Component Models: Theory and Applications*, pages 123–151. Cambridge University Press, Cambridge.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.
- Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partially non-Gaussian state space models. *Journal of Econometrics*, 154:85–100.
- Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, pages 165–200. Oxford University Press, Oxford (UK).
- Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26:216–230.
- Geweke, J. and Jiang, Y. (2011). Inference and prediction in a multiple-structural-break model. *Journal of Econometrics*, 163:172–185.
- Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, 138:252–291.
- Geweke, J. and Tanizaki, H. (1999). On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state space models. *Communications in Statistics, Part B – Simulation and Computation*, 28:867–894.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.
- Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold and quantile weighted proper scoring rules. *Journal of Business & Economic Statistics*, 29:411–422.
- Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Ser. B*, pages 107–114.
- Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5:171–188.
- Hörmann, W. and Leydold, J. (2014). Generating generalized inverse gaussian random variates. *Statistics and Computing*, 24:547–557.

- Hörmann, W. and Leydold, J. (2015). GIGrv: Random variate generator for the gig distribution. R package version 0.4, url: <http://CRAN.R-project.org/package=GIGrv>.
- Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, 12:371–417.
- Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793.
- Kastner, G. (2016). Dealing with stochastic volatility in time series using the R package stochvol. *Journal of Statistical Software*, 69:1–30.
- Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics and Data Analysis*, 76:408–423.
- Kastner, G., Frühwirth-Schnatter, S., and Lopes, H. F. (2016). Efficient Bayesian inference for multivariate factor stochastic volatility models. *Submitted*.
- Leydold, J. and Hörmann, W. (2011). Generating generalized inverse gaussian random variates by fast inversion. *Computational Statistics & Data Analysis*, 55(1):213–217.
- Lopes, H. F., McCulloch, R. E., and Tsay, R. S. (2015). Parsimony inducing priors for large scale state-space models. *Discussion Paper*.
- McCausland, W. J., Miller, S., and Pelletier, D. (2011). Simulation smoothing for state space models: A computational efficiency analysis. *Computational Statistics and Data Analysis*, 55:199 – 212.
- Nakajima, J. (2011). Time-varying parameter VAR model with stochastic volatility: An overview of methodology and empirical applications. *Monetary and Economic Studies*, 29:107–142.
- Nakajima, J. and West, M. (2013). Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31:151–164.
- O’Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4:85–118.
- Papaspiliopoulos, O., Roberts, G., and Sköld, M. (2007). A general framework for the parameterization of hierarchical models. *Statistical Science*, 22:59–73.
- Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686.
- Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Springer, New York.

- Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, pages 501–538. Oxford University Press, Oxford (UK).
- Primiceri, G. (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72:821–852.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London.
- Simpson, M., Niemi, J., and Roy, V. (2015). Interweaving Markov chain Monte Carlo strategies for efficient estimation of dynamic linear models. *Journal of Computational and Graphical Statistics*, (accepted).
- Sims, C. A. (2001). [Evolving post-world war II US inflation dynamics]: Comment. *NBER Macroeconomics Annual*, 16:373–379.
- Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, 58:267–288.
- Villani, M., Kohn, R., and Giordani, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics*, 153:155–173.
- West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer, New York, 2 edition.
- Yu, Y. and Meng, X.-L. (2011). To center or not to center: that is not the question - an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.
- Zhao, Z. Y., Xie, M., and West, M. (2016). Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32:311–332.

A.1 Computational Details

A.1.1 Details on the MCMC scheme in Algorithm 1

A.1.1.1 Step (a): Sampling the latent states

Step (a) of our Gibbs sampler is to sample the latent states conditional on known parameters using either *Forward Filtering Backward Sampling* (FFBS), as discussed in Frühwirth-Schnatter (1994) and Carter and Kohn (1994), or the faster alternative known as *all without a loop* (AWOL), discussed in McCausland et al. (2011) and Kastner and Frühwirth-Schnatter (2014). Subsequently, we provide details how the AWOL algorithm is implemented for TVP models.

The algorithm is implemented for a slight modification of the non-centered state space model (6) and (7), given by:

$$\begin{aligned}\tilde{\beta}_{jt} &= \tilde{\beta}_{j,t-1} + \tilde{\omega}_{jt}, \quad \tilde{\omega}_{jt} \sim \mathcal{N}(0, 1), \\ y_t^* &= \mathbf{F}_t \tilde{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2),\end{aligned}$$

with outcome $y_t^* = y_t - \mathbf{x}_t \boldsymbol{\beta}$ and $\mathbf{F}_t = \mathbf{x}_t \text{Diag}(\sqrt{\theta_1}, \dots, \sqrt{\theta_d})$ for $t = 1, \dots, T$.

Conditional on all other variables, the joint density for the state vector $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_T)$ is multivariate normal. This distribution can be written in terms of the tri-diagonal precision matrix $\boldsymbol{\Omega}$ and the covector \mathbf{c} , see also Rue and Held (2005):

$$\tilde{\boldsymbol{\beta}} \sim \mathcal{N}_{(T+1) \cdot d}(\boldsymbol{\Omega}^{-1} \mathbf{c}, \boldsymbol{\Omega}^{-1}), \quad (\text{A.1})$$

where:

$$\boldsymbol{\Omega} \equiv \begin{bmatrix} \boldsymbol{\Omega}_{00} & \boldsymbol{\Omega}_{01} & 0 & & & \\ \boldsymbol{\Omega}'_{10} & \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} & 0 & & \\ 0 & \boldsymbol{\Omega}'_{12} & \boldsymbol{\Omega}_{22} & \boldsymbol{\Omega}_{23} & \ddots & \vdots \\ & 0 & \boldsymbol{\Omega}'_{23} & \ddots & \ddots & 0 \\ & \vdots & \ddots & \ddots & \boldsymbol{\Omega}_{T-1,T-1} & \boldsymbol{\Omega}_{T-1,T} \\ & 0 & \dots & 0 & \boldsymbol{\Omega}'_{T-1,T} & \boldsymbol{\Omega}_{TT} \end{bmatrix}, \quad \mathbf{c} \equiv \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_T \end{bmatrix}.$$

In this representation, each submatrix $\boldsymbol{\Omega}_{ts}$ is a matrix of dimension $d \times d$, defined as

$$\begin{aligned}\boldsymbol{\Omega}_{00} &\equiv \text{Diag}(1/P_{0,11} \cdots 1/P_{0,dd}) + \mathbf{I}_d, \\ \boldsymbol{\Omega}_{tt} &\equiv \mathbf{F}_t' \mathbf{F}_t / \sigma_t^2 + 2\mathbf{I}_d, \quad t = 1, \dots, T-1, \\ \boldsymbol{\Omega}_{TT} &\equiv \mathbf{F}_T' \mathbf{F}_T / \sigma_T^2 + \mathbf{I}_d, \\ \boldsymbol{\Omega}_{t,t+1} &\equiv -\mathbf{I}_d, \quad t = 0, \dots, T-1,\end{aligned}$$

whereas each \mathbf{c}_t is a vector of dimension $d \times 1$, defined as

$$\mathbf{c}_0 \equiv \mathbf{0}, \quad \mathbf{c}_t \equiv (\mathbf{F}'_t / \sigma_t^2) y_t^*, \quad t = 1, \dots, T.$$

The specific structure of $\mathbf{\Omega}$ allows sampling the state process $\tilde{\beta}_0, \dots, \tilde{\beta}_T$, *all without a loop* from the posterior $\tilde{\beta} \sim \mathcal{N}_{(T+1) \cdot d}(\mathbf{\Omega}^{-1} \mathbf{c}, \mathbf{\Omega}^{-1})$. Due to the band structure of $\mathbf{\Omega}$, calculating the Cholesky decomposition $\mathbf{\Omega} = \mathbf{L}\mathbf{L}'$ is computationally inexpensive. Based on a draw $\epsilon \sim \mathcal{N}_{(T+1) \cdot d}(\mathbf{0}, \mathbf{I})$, we solve $\mathbf{L}\mathbf{a} = \mathbf{c}$ for \mathbf{a} and $\mathbf{L}'\tilde{\beta} = \mathbf{a} + \epsilon$ for $\tilde{\beta}$ by using back-band substitution instead of actually calculating \mathbf{L}^{-1} . Further details on this method can be found in McCausland et al. (2011).

A.1.1.2 Step (b): Sampling the constant coefficient β_j and the square root of the error variance $\sqrt{\theta_j}$

Conditional on the state process $\tilde{\beta} = (\tilde{\beta}_0, \dots, \tilde{\beta}_T)$, the observation equation (7) of the non-centered state space model defines an expanded regression model:

$$y_t = \mathbf{z}_t \boldsymbol{\alpha} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \quad (\text{A.2})$$

with regression coefficient $\boldsymbol{\alpha} = (\beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d})'$ and covariate vector $\mathbf{z}_t = (\mathbf{x}_t, x_{t1}\tilde{\beta}_{1t}, \dots, x_{td}\tilde{\beta}_{dt})$. Under the conjugate prior $\boldsymbol{\alpha} \sim \mathcal{N}_{2d}(\mathbf{a}_0, \mathbf{A}_0)$, where $\mathbf{a}_0 = \mathbf{0}$ and $\mathbf{A}_0 = \text{Diag}(\tau_1^2, \dots, \tau_d^2, \xi_1^2, \dots, \xi_d^2)$, it follows that the conditional posterior distribution $p(\boldsymbol{\alpha} | \tilde{\beta}, \sigma_1^2, \dots, \sigma_T^2, \boldsymbol{\tau}, \boldsymbol{\xi}, \mathbf{y})$ is a multivariate normal distribution,

$$\boldsymbol{\alpha} | \tilde{\beta}, \sigma_1^2, \dots, \sigma_T^2, \boldsymbol{\tau}, \boldsymbol{\xi}, \mathbf{y} \sim \mathcal{N}_{2d}(\mathbf{a}_T, \mathbf{A}_T), \quad (\text{A.3})$$

with

$$\mathbf{a}_T = \mathbf{A}_T (\tilde{\mathbf{W}}\mathbf{y} + \mathbf{A}_0^{-1} \mathbf{a}_0), \quad \mathbf{A}_T = (\tilde{\mathbf{W}}\mathbf{W} + \mathbf{A}_0^{-1})^{-1},$$

where $\mathbf{y} = (y_1, \dots, y_T)'$ and \mathbf{W} is a $(T \times 2d)$ regressor matrix with the t -th row being equal to \mathbf{z}_t and $\tilde{\mathbf{W}} = \mathbf{W}' \text{Diag}(1/\sigma_1^2, \dots, 1/\sigma_T^2)$.

In a shrinkage framework, some of the variances τ_j^2 and ξ_j^2 may be close to 0, leading to conditional priors $p(\beta_j | \tau_j^2)$ and $p(\sqrt{\theta_j} | \xi_j^2)$ with huge prior information equal to $1/\tau_j^2$ and $1/\xi_j^2$. In order to overcome numerical difficulties with such very informative priors, we make use of the following alternative computation of the posterior covariance matrix \mathbf{A}_T :

$$\begin{aligned} \mathbf{A}_T &= \mathbf{A}_0^{1/2} \mathbf{A}_T^* \mathbf{A}_0^{1/2}, \\ \mathbf{A}_0^{1/2} &= \text{Diag}(\tau_1, \dots, \tau_d, \xi_1, \dots, \xi_d), \quad \mathbf{A}_T^* = \left(\mathbf{A}_0^{1/2} \tilde{\mathbf{W}} \mathbf{W} \mathbf{A}_0^{1/2} + \mathbf{I}_{2d} \right)^{-1}. \end{aligned} \quad (\text{A.4})$$

A.1.1.3 Step (c): Sampling the prior variances

For the normal-gamma hierarchical priors it follows that the conditionally normal prior $\beta_j|\tau_j^2$ ($\sqrt{\theta_j}|\xi_j^2$) leads to a posterior for the variance $\tau_j^2|\beta_j$ ($\xi_j^2|\theta_j$), where the likelihood is the kernel of an inverted gamma density in τ_j^2 (ξ_j^2). In combination with the gamma prior $\tau_j^2 \sim \mathcal{G}(a^\tau, a^\tau \lambda^2/2)$ ($\xi_j^2 \sim \mathcal{G}(a^\xi, a^\xi \kappa^2/2)$), this leads to a generalized inverse Gaussian distribution for $\tau_j^2|\beta_j$ ($\xi_j^2|\theta_j$):

$$\xi_j^2|\theta_j, a^\xi, \kappa^2 \sim \mathcal{GIG}\left(a^\xi - 1/2, a^\xi \kappa^2, \theta_j\right), \quad (\text{A.5})$$

$$\tau_j^2|\beta_j, a^\tau, \lambda^2 \sim \mathcal{GIG}\left(a^\tau - 1/2, a^\tau \lambda^2, \beta_j\right). \quad (\text{A.6})$$

The generalized inverse Gaussian distribution, $Y \sim \mathcal{GIG}(p, a, b)$, $a > 0, b > 0$ is a three parameter family with support on $y \in \mathbb{R}^+$. The density is given by

$$p(y) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} y^{p-1} e^{-(a/2)y} e^{-b/(2y)}, \quad (\text{A.7})$$

and $K_p(z)$ is the modified Bessel function of the second kind. The moments $\mu_k = \mathbb{E}(Y^k)$ are given as

$$\mu_k = \frac{K_{p+k}(\sqrt{ab})}{K_p(\sqrt{ab})} \left(\sqrt{\frac{b}{a}} \right)^k, \quad k \in \mathbb{R}. \quad (\text{A.8})$$

Hörmann and Leydold (2014) propose a new generation method for the cases where $p < 1, \sqrt{ab} < 0.5$, which is especially useful in the time-varying parameter case. A very stable generator is implemented in the R-package GIGrvrg (Leydold and Hörmann, 2011).

As we have specified κ^2 and λ^2 to be unknown with hyper-priors $\kappa^2 \sim \mathcal{G}(d_1, d_2)$ and $\lambda^2 \sim \mathcal{G}(e_1, e_2)$, we need to sample these parameters from the corresponding conditional posteriors:

$$\kappa^2|a^\xi, d_1, d_2, \xi_1^2, \dots, \xi_d^2 \sim \mathcal{G}\left(d_1 + a^\xi d, d_2 + \frac{\overline{\xi^2}}{2} a^\xi d\right), \quad (\text{A.9})$$

$$\lambda^2|a^\tau, e_1, e_2, \tau_1^2, \dots, \tau_d^2 \sim \mathcal{G}\left(e_1 + a^\tau d, e_2 + \frac{\overline{\tau^2}}{2} a^\tau d\right), \quad (\text{A.10})$$

where $\overline{\xi^2}$ and $\overline{\tau^2}$ are the averages of the variances in the shrinkage priors:

$$\overline{\xi^2} = \frac{1}{d} \sum_{j=1}^d \xi_j^2, \quad \overline{\tau^2} = \frac{1}{d} \sum_{j=1}^d \tau_j^2.$$

If $\overline{\xi^2}$ is small, i.e. a lot of sparsity is present, then the posterior expectation of κ^2 will be proportional to $1/d_2$. Hence, smaller values of d_2 encourage stronger prior shrinkage.

A.1.2 Using the Kalman filter for prediction

It is well-known that the Kalman filter can be applied to derive the predictive density in a state space model for known model parameters. We exploit this procedure, to derive the predictive density $p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\beta}, \mathbf{Q}, \sigma_t^2)$ for known values of the parameters $\boldsymbol{\beta}$, \mathbf{Q} , and σ_t^2 , given observations $\mathbf{y}^{t-1} = (y_1, \dots, y_{t-1})$. In the following, details are provided for the non-centered parameterization of the TVP model, based on rewriting the observation equation (7) as:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_t &= \tilde{\boldsymbol{\beta}}_{t-1} + \tilde{\boldsymbol{\omega}}_t, & \tilde{\boldsymbol{\omega}}_t &\sim \mathcal{N}_d(0, \mathbf{I}_d), \\ y_t &= \mathbf{x}_t\boldsymbol{\beta} + \mathbf{F}_t\tilde{\boldsymbol{\beta}}_t + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, \sigma_t^2),\end{aligned}$$

where $\mathbf{F}_t = \mathbf{x}_t \text{Diag}(\sqrt{\theta_1}, \dots, \sqrt{\theta_d})$. Starting from $\tilde{\boldsymbol{\beta}}_0 \sim \mathcal{N}_d(\mathbf{m}_0, \mathbf{C}_0)$ with $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{C}_0 = \mathbf{P}_0 = \text{Diag}(P_{0,11} \cdots P_{0,dd})$, the following three steps are repeated for $t = 1, \dots, T-1$:⁸

- (a) A *propagation step* to determine the one-step ahead predictive density $p(\tilde{\boldsymbol{\beta}}_t|\mathbf{y}^{t-1})$:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_t|\mathbf{y}^{t-1} &\sim \mathcal{N}_d(\mathbf{m}_{t-1}, \mathbf{R}_t), \\ \mathbf{R}_t &= \mathbf{C}_{t-1} + \mathbf{I}_d.\end{aligned}$$

- (b) A *prediction step* to determine the predictive density $p(y_t|\mathbf{y}^{t-1})$:

$$\begin{aligned}y_t|\mathbf{y}^{t-1} &\sim \mathcal{N}(\hat{y}_t, S_t), \\ \hat{y}_t &= \mathbf{F}_t\mathbf{m}_{t-1} + \mathbf{x}_t\boldsymbol{\beta}, & S_t &= \mathbf{F}_t\mathbf{R}_t\mathbf{F}_t' + \sigma_t^2.\end{aligned}$$

- (c) A *correction step* to determine the filter density $p(\tilde{\boldsymbol{\beta}}_t|\mathbf{y}^t)$:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_t|\mathbf{y}^t &\sim \mathcal{N}_d(\mathbf{m}_t, \mathbf{C}_t), \\ \mathbf{m}_t &= \mathbf{m}_{t-1} + \mathbf{K}_t(y_t - \hat{y}_t), & \mathbf{K}_t &= \mathbf{R}_t\mathbf{F}_t'S_t^{-1}, & \mathbf{C}_t &= (\mathbf{I}_d - \mathbf{K}_t\mathbf{F}_t)\mathbf{R}_t.\end{aligned}$$

Note that prediction could equally well be performed in the centered parametrization of the TVP model.

⁸To simplify the notation, dependence on the model parameters $\boldsymbol{\beta}$, \mathbf{Q} , and $\sigma_1^2, \dots, \sigma_T^2$ is not made explicit.

A.2 Data description

A.2.1 ECB Data

The ECB data we have used can be retrieved freely from the ECB datawarehouse⁹. Our data are monthly and range from February 1994 until November 2010. The original time series have been transformed (log differences, differences) and each of the time series has been standardised. The core inflation rate has been transformed to have variance one. Graphs of the original time series before transformations have been applied are shown in Figure A.1. In the following the times series are listed in detail:

1. I-1MO one-month Euribor (Euro interbank offered rate).
2. I-1YR: one-year Euribor (Euro interbank offered rate).
3. SENT: Percentage change in economic sentiment indicator.
4. STOCK-1: Percentage change in equity index - Dow Jones, Euro Stoxx, Economic sector index financial.
5. STOCK-2: Percentage change in equity index - Dow Jones Euro Stoxx 50 index.
6. EXRATE: Percentage change in ECB real effective exchange rate (CPI deflated, broad group of currencies against euro).
7. IP: Percentage change in industrial production index.
8. LOANS: Percentage change in loans (total maturity, all currencies combined).
9. M3: Annual percentage change in monetary aggregate M3.
10. CAR: Registrations of new passenger cars.
11. OIL: Percentage change in oil price (brent crude, 1-month forward).
12. ORDER: Change in order-book levels.
13. UNEMP: Standardised unemployment rate (all ages, male & female).
14. HICP: harmonized index of consumer prices

A.2.2 DAX Data

The DAX - Deutscher Aktienindex (German stock index) is a blue chip stock market index consisting of the 30 major German companies trading on the Frankfurt Stock Exchange, see Table A.1. Our data set spans roughly 2500 daily stock returns from September 4th 2001 until August 31st, 2011. One of those companies is excluded from our case study, as it was not part of the DAX for the whole time span.

⁹For details and data look at <http://sdw.ecb.europa.eu/>.

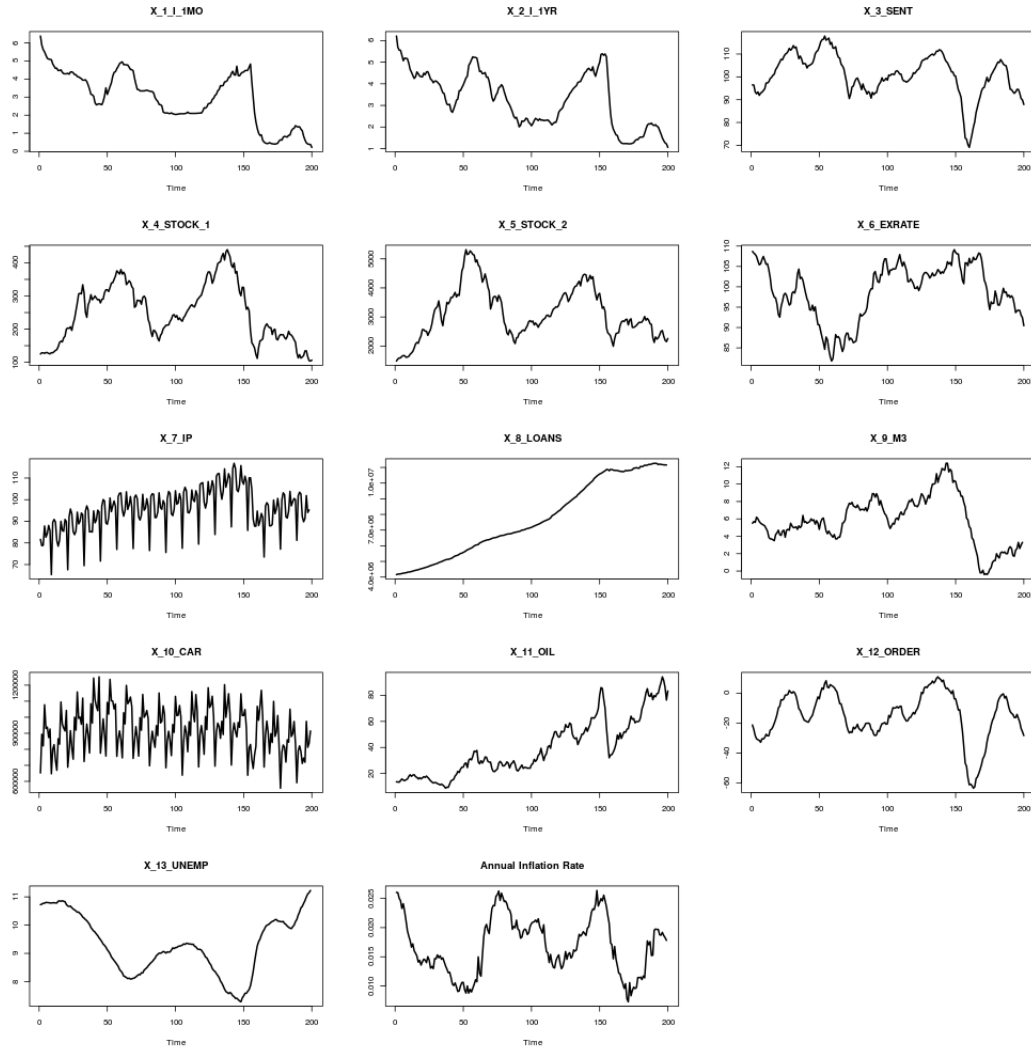


Figure A.1: ECB data. Original time series.

Table A.1: DAX data. Data description.

<i>i</i>	Symbol	Name	<i>i</i>	Symbol	Name
1	ADS	Adidas	16	FRE	Fresenius
2	ALV	Allianz	17	FME	Fresenius Medical Care
3	BAS	BASF	18	HEI	HeidelbergCement
4	BAYN	Bayer	19	HEN3	Henkel
5	BEI	Beiersdorf	20	IFX	Infineon Technologies
6	BMW	BMW	21	SDF	K + S
7	CBK	Commerzbank	22	LIN	Linde
8	CON	Continental	23	MRK	Merck
9	DAI	Daimler	24	MUV2	Munich Re
10	DBK	Deutsche Bank	25	RWE	RWE
11	DB1	Deutsche Börse	26	SAP	SAP
12	LHA	Deutsche Lufthansa	27	SIE	Siemens
13	DPW	Deutsche Post	28	TKA	ThyssenKrupp
14	DTE	Deutsche Telekom	29	VOW3	Volkswagen Group
15	EOAN	E.ON			

A.3 Further results

	Name	Mean	Stdev.	Median	2.5% quantil	97.5% quantil
β_1	Intercept	3.8632E+00	7.1818E-01	3.8863E+00	2.3808E+00	5.2126E+00
β_2	lag1	-7.4362E-04	3.0279E-02	-9.6009E-06	-6.4674E-02	6.5412E-02
β_3	lag2	-1.2718E-02	3.3511E-02	-1.7024E-03	-9.9431E-02	3.8729E-02
β_4	lag3	-1.8394E-02	3.3205E-02	-6.2266E-03	-1.0120E-01	2.9296E-02
β_5	lag4	-1.5157E-03	2.5692E-02	-1.3518E-06	-6.2836E-02	5.1550E-02
β_6	lag5	-8.6116E-03	2.8575E-02	-1.0041E-03	-7.9285E-02	4.2273E-02
β_7	lag6	2.8570E-02	3.7900E-02	1.8229E-02	-2.2479E-02	1.1768E-01
β_8	lag7	3.2783E-03	2.7267E-02	9.7368E-05	-5.4998E-02	6.5993E-02
β_9	lag8	-2.2960E-02	3.9891E-02	-7.5587E-03	-1.2658E-01	3.0400E-02
β_{10}	lag9	-6.4052E-03	3.6299E-02	-1.6396E-04	-9.9082E-02	5.9776E-02
β_{11}	lag10	-3.5588E-02	4.7422E-02	-2.0986E-02	-1.5176E-01	2.5221E-02
β_{12}	lag11	-2.7056E-02	4.3923E-02	-9.8708E-03	-1.4058E-01	2.8533E-02
β_{13}	lag12	-8.6966E-02	6.9579E-02	-8.8094E-02	-2.2296E-01	2.3730E-02
β_{14}	x1_I1MO	4.4239E-01	2.7247E-01	4.4241E-01	-1.1595E-02	9.9787E-01
β_{15}	x2_I1YR	4.8565E-02	1.7514E-01	2.1806E-03	-2.4179E-01	5.3345E-01
β_{16}	x3_sent	-8.6713E-03	2.6065E-02	-1.5612E-03	-6.8655E-02	3.5499E-02
β_{17}	x4_stock1	1.5780E-02	3.1857E-02	4.5085E-03	-2.9750E-02	9.5954E-02
β_{18}	x5_stock2	-1.7107E-04	2.7899E-02	9.4073E-08	-6.2637E-02	5.8218E-02
β_{19}	x6_exrate	3.6063E-03	1.8920E-02	3.8224E-04	-3.4840E-02	4.4487E-02
β_{20}	x7_ip	2.8744E-03	1.3509E-02	2.0712E-04	-2.2615E-02	3.4475E-02
β_{21}	x8_loans	-2.6774E-03	2.1952E-02	-7.2792E-05	-5.1865E-02	3.9251E-02
β_{22}	x9_M3	-2.8925E-02	1.7184E-01	-4.0903E-06	-5.1438E-01	2.4440E-01
β_{23}	x10_car	4.7635E-03	2.5995E-02	4.2945E-04	-4.9982E-02	6.2336E-02
β_{24}	x11_oil	-6.8976E-03	1.7183E-02	-1.3893E-03	-4.8858E-02	1.9612E-02
β_{25}	x12_order	5.6852E-03	2.3943E-02	2.6091E-04	-3.2064E-02	6.4826E-02
β_{26}	x13_unemp	2.5488E-02	2.5509E-01	1.0083E-05	-4.6216E-01	7.2208E-01
β_{27}	d1	3.7838E-03	5.8718E-02	1.2084E-06	-1.1331E-01	1.4674E-01
β_{28}	d2	-7.8427E-04	5.3182E-02	7.4369E-08	-1.2473E-01	1.1574E-01
β_{29}	d3	2.2424E-02	7.1359E-02	3.8909E-03	-1.1711E-01	1.8728E-01
β_{30}	d4	3.6066E-02	6.3320E-02	1.1637E-02	-5.4160E-02	1.9491E-01
β_{31}	d5	1.4198E-02	5.2411E-02	1.1188E-03	-8.2077E-02	1.4631E-01
β_{32}	d6	-2.0833E-02	5.4098E-02	-2.9700E-03	-1.5712E-01	7.2195E-02
β_{33}	d7	-4.6528E-02	7.1441E-02	-2.0850E-02	-2.1933E-01	5.3918E-02
β_{34}	d8	-5.9570E-03	4.8426E-02	-1.1442E-04	-1.2326E-01	9.5339E-02
β_{35}	d9	-2.4944E-03	4.7404E-02	-3.0573E-06	-1.1378E-01	9.8459E-02
β_{36}	d10	-5.4952E-03	4.8600E-02	-6.8190E-05	-1.2365E-01	9.4591E-02
β_{37}	d11	8.8116E-02	9.3950E-02	7.5676E-02	-4.8433E-02	2.8535E-01
$ \sqrt{\theta_1} $	Intercept	1.5711E-01	2.5018E-02	1.5809E-01	1.0568E-01	2.0320E-01
$ \sqrt{\theta_2} $	lag1	2.2215E-03	3.1962E-03	9.4102E-04	2.0302E-07	1.1405E-02

Continued on next page

Table A.2 – *Continued from previous page*

	Name	Mean	Stdev.	Median	2.5% quantil	97.5% quantil
$ \sqrt{\theta_3} $	lag2	1.8972E-03	2.8663E-03	7.9946E-04	1.3573E-07	9.9332E-03
$ \sqrt{\theta_4} $	lag3	1.8101E-03	2.5575E-03	8.0924E-04	1.0047E-07	8.9229E-03
$ \sqrt{\theta_5} $	lag4	1.5493E-03	2.1936E-03	6.7867E-04	8.7567E-08	7.6548E-03
$ \sqrt{\theta_6} $	lag5	1.9234E-03	2.7252E-03	8.3776E-04	1.1669E-07	9.5874E-03
$ \sqrt{\theta_7} $	lag6	1.7706E-03	2.4146E-03	8.0243E-04	7.4876E-08	8.5274E-03
$ \sqrt{\theta_8} $	lag7	1.4714E-03	2.0760E-03	6.5023E-04	1.6423E-08	7.3677E-03
$ \sqrt{\theta_9} $	lag8	1.9251E-03	2.8745E-03	7.9588E-04	7.8992E-08	1.0173E-02
$ \sqrt{\theta_{10}} $	lag9	2.6633E-03	3.9409E-03	1.1137E-03	3.5889E-07	1.4226E-02
$ \sqrt{\theta_{11}} $	lag10	2.1613E-03	2.8942E-03	1.0354E-03	2.7814E-07	1.0190E-02
$ \sqrt{\theta_{12}} $	lag11	2.0186E-03	2.8059E-03	9.0706E-04	1.2576E-07	9.8949E-03
$ \sqrt{\theta_{13}} $	lag12	3.9779E-03	4.9794E-03	1.9718E-03	2.6016E-07	1.7477E-02
$ \sqrt{\theta_{14}} $	x1.L1MO	2.4478E-02	2.6748E-02	1.3967E-02	5.8272E-06	9.0774E-02
$ \sqrt{\theta_{15}} $	x2.L1YR	1.9668E-02	2.1759E-02	1.1934E-02	4.6017E-06	7.6396E-02
$ \sqrt{\theta_{16}} $	x3_sent	2.6433E-03	4.2631E-03	1.0300E-03	2.5725E-07	1.4404E-02
$ \sqrt{\theta_{17}} $	x4_stock1	2.2995E-03	3.3794E-03	9.4762E-04	7.4584E-08	1.2027E-02
$ \sqrt{\theta_{18}} $	x5_stock2	2.5676E-03	3.8623E-03	1.0483E-03	1.2417E-07	1.3659E-02
$ \sqrt{\theta_{19}} $	x6_exrate	2.3525E-03	3.4434E-03	1.0091E-03	1.2544E-07	1.2187E-02
$ \sqrt{\theta_{20}} $	x7_ip	1.3423E-03	1.9714E-03	5.6948E-04	1.3191E-07	6.9567E-03
$ \sqrt{\theta_{21}} $	x8_loans	2.7280E-03	4.2800E-03	1.0379E-03	2.2741E-07	1.5105E-02
$ \sqrt{\theta_{22}} $	x9_M3	2.9196E-02	2.3609E-02	2.5564E-02	2.3118E-05	8.1521E-02
$ \sqrt{\theta_{23}} $	x10_car	4.9918E-03	6.0645E-03	2.8384E-03	9.5520E-07	2.1199E-02
$ \sqrt{\theta_{24}} $	x11_oil	1.6200E-03	2.3494E-03	6.8330E-04	6.6702E-08	8.2332E-03
$ \sqrt{\theta_{25}} $	x12_order	2.2634E-03	3.4616E-03	8.8449E-04	1.0748E-07	1.2245E-02
$ \sqrt{\theta_{26}} $	x13_unemp	4.9005E-02	3.4878E-02	5.1310E-02	3.4136E-05	1.1262E-01
$ \sqrt{\theta_{27}} $	d1	6.0722E-03	7.8571E-03	2.8768E-03	4.5119E-07	2.7619E-02
$ \sqrt{\theta_{28}} $	d2	8.0497E-03	9.5195E-03	4.4726E-03	1.3434E-06	3.3239E-02
$ \sqrt{\theta_{29}} $	d3	1.1010E-02	1.2101E-02	7.2259E-03	2.2084E-06	4.2559E-02
$ \sqrt{\theta_{30}} $	d4	4.8371E-03	6.7577E-03	2.1482E-03	3.6042E-07	2.3544E-02
$ \sqrt{\theta_{31}} $	d5	4.6158E-03	6.0828E-03	2.1830E-03	4.5843E-07	2.1337E-02
$ \sqrt{\theta_{32}} $	d6	4.3897E-03	5.8349E-03	2.0446E-03	6.2302E-07	2.0687E-02
$ \sqrt{\theta_{33}} $	d7	6.3108E-03	7.9964E-03	3.1448E-03	7.8058E-07	2.7908E-02
$ \sqrt{\theta_{34}} $	d8	4.3322E-03	5.9656E-03	1.9354E-03	5.0841E-07	2.1238E-02
$ \sqrt{\theta_{35}} $	d9	5.3348E-03	6.9779E-03	2.5517E-03	3.7276E-07	2.4663E-02
$ \sqrt{\theta_{36}} $	d10	4.8326E-03	6.6309E-03	2.2330E-03	6.5966E-07	2.3436E-02
$ \sqrt{\theta_{37}} $	d11	9.5025E-03	1.0321E-02	6.2378E-03	1.7492E-06	3.5477E-02
σ^2		8.1435E-03	3.3801E-03	7.6922E-03	2.8696E-03	1.5952E-02

Table A.2: ECB data. Posterior summary statistics under a shrinkage prior with $a^\tau = a^\xi = 0.2$.

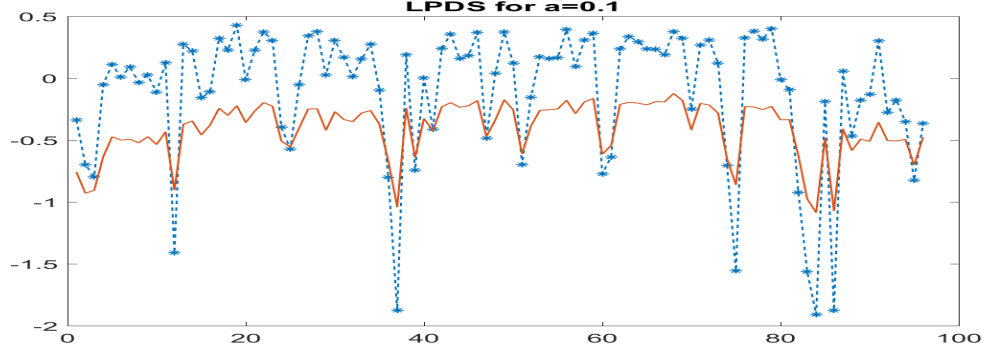


Figure A.2: ECB data. Comparing approximations of the log predictive densities scores LPDS_{t+1}^* for $a^\tau = a^\xi = 0.1$ obtained from the optimal Kalman mixture approximation (full line) and the naive Gaussian mixture approximation (dashed line with \star).

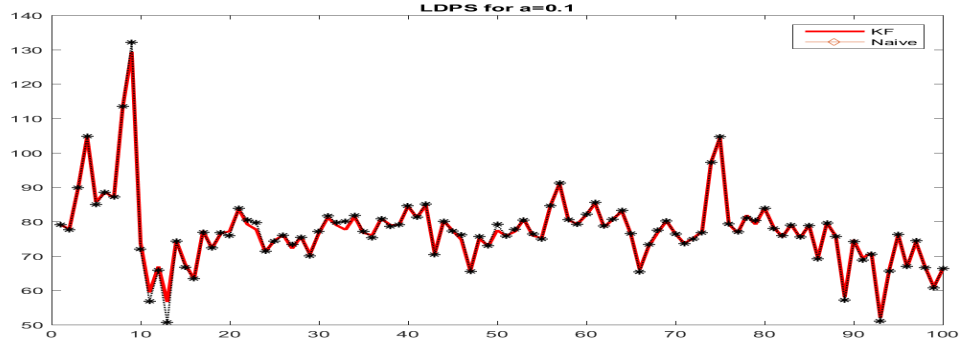


Figure A.3: DAX data. Comparing approximations of the overall log predictive densities scores LPDS_{t+1}^* for $a^\tau = a^\xi = 0.1$ obtained from the optimal Kalman mixture approximation (full line) and the naive Gaussian mixture approximation (dashed line with \star).

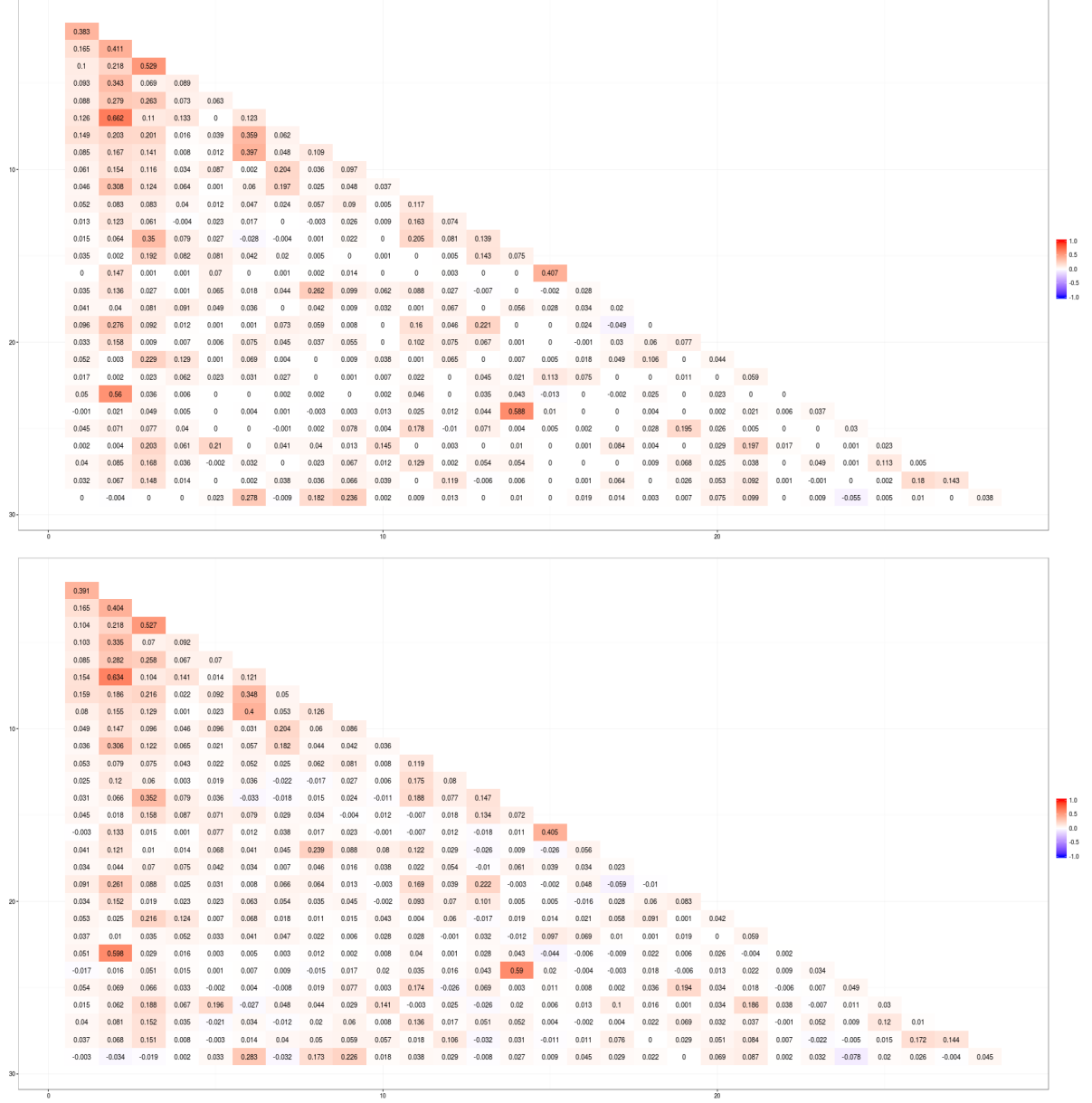


Figure A.4: DAX data. Heat plot of the posterior median of \mathbf{B}_t at $t = 1150$ for $a^\tau = a^\xi = 0.1$ (top) and $a^\tau = a^\xi = 1$ (bottom).